Call: HORIZON-HLTH-2022-TOOL-11
Topic: HORIZON-HLTH-2022-TOOL-11-02
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101095435

Deliverable No. 4.6

Multimodal Synthetic Data Generator

Contractual Submission Date: 30/09/2025

Actual Submission Date: 30/09/2025

Responsible partner: P01-UM

Funded by
the European Union

Grant agreement no.	101095435
Project full title	REALM - Real-world-data Enabled Assessment for heaLth regulatory decision-Making

Deliverable number	**D4.6**
Deliverable title	**Multimodal Synthetic Data Generator**
Type[1]	R — Document, report
Dissemination level[2]	PU - Public
Work package number	WP4
Work package leader	P1-UM
Author(s)	Mahmoud Ibrahim (UM, VITO) Chang Sun (UM) Bart Elen (VITO) Janaki Raman Rangarajan (Vphi)
Keywords	Synthetic Data Generation; Generative Models; Digital Twins; Privacy Preserving;

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA).

Neither the European Union nor the granting authority can be held responsible for them.

Document History

Version	Date	Description
V0.1	03/09/2025	Initial Draft for internal review
V0.2	21/09/2025	Internal feedback submitted
V0.3	25/09/2025	Internal feedback addressed
V1	30/09/2025	Final Version for submission

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)
PU: Public, fully open, e.g. web
SEN: Sensitive, limited under conditions of the Grant Agreement

Table of Contents

List of Abbreviations and definitions

A

- **AE** - Autoencoder
- **AEKL** - Autoencoder Kullback-Leibler
- **AI** - Artificial Intelligence
- **AML** - Acute Myeloid Leukemia
- **AP** - Anteroposterior
- **AUC** - Area Under the Curve
- **AUC-PR** - Area Under the Curve - Precision-Recall
- **AUC-ROC** - Area Under the Curve - Receiver Operating Characteristic

B

- **BCE** - Binary Cross-Entropy
- **BLEU** - Bilingual Evaluation Understudy
- **BMI** - Body Mass Index
- **BPE** - Byte Pair Encoding
- **BRNN** - Bidirectional Recurrent Neural Network

C

- **CC** - Cross-correlation
- **cGAN** - Conditional Generative Adversarial Network
- **CI** - Confidence Interval
- **CLIP** - Contrastive Language-Image Pre-training
- **CNR** - Contrast Noise Ratio
- **CNN** - Convolutional Neural Network
- **COPD** - Chronic Obstructive Pulmonary Disease
- **CSA** - Coordinate & Support Action
- **CT** - Computed Tomography
- **CTGAN** - Conditional Tabular Generative Adversarial Network
- **CXR** - Chest X-Ray

D

- **DCGAN** - Deep Convolutional Generative Adversarial Network
- **DDPM** - Denoising Diffusion Probabilistic Model

- **DL** - Deep Learning
- **DM** - Diffusion Model
- **DMP** - Data Management Plan
- **DP** - Differential Privacy
- **DP-CGAN** - Differentially Private Conditional Generative Adversarial Network
- **DP-SGD** - Differentially Private Stochastic Gradient Descent
- **DPM** - Diffusion Probabilistic Model
- **DS** - Diversity Score
- **DTW** - Dynamic Time Warping

E

- **ECG** - Electrocardiogram
- **EDITH** - Ecosystem of Digital Twins in Healthcare
- **EEG** - Electroencephalogram
- **EHR** - Electronic Health Record

F

- **FAIR** - Findable, Accessible, Interoperable, Reusable
- **FDS** - Feature Distribution Similarity
- **FID** - Fréchet Inception Distance
- **FSIM** - Feature Similarity Index

G

- **GAN** - Generative Adversarial Network
- **GDPR** - General Data Protection Regulation
- **GPU** - Graphics Processing Unit
- **GRU** - Gated Recurrent Unit

H

- **HOOM** - (Human Ontology context, specific definition not provided)
- **HPO** - Human Phenotype Ontology

I

- **ICD** - International Classification of Diseases
- **ICU** - Intensive Care Unit
- **IS** - Inception Score

J

- **JSON** - JavaScript Object Notation

K

- **KID** - Kernel Inception Distance
- **KLD** - Kullback-Leibler Divergence

L

- **LDM** - Latent Diffusion Model
- **LLM** - Large Language Model
- **LOS** - Length of Stay
- **LPIPS** - Learned Perceptual Image Patch Similarity
- **LSGAN** - Least Squares Generative Adversarial Network
- **LSTM** - Long Short-term Memory

M

- **MCC** - Matthews Correlation Coefficient
- **MMD** - Maximum Mean Discrepancy
- **MRI** - Magnetic Resonance Imaging
- **MS-SSIM** - Multi-Scale Structural Similarity Index Measure
- **MSE** - Mean Squared Error

N

- **NER** - Named Entity Recognition
- **NLP** - Natural Language Processing
- **NNAA** - Nearest Neighbour Adversarial Accuracy
- **NRMSE** - Normalized Root Mean Squared Error
- **O**
- **ORDO** - Orphanet Rare Disease Ontology

P

- **PA** - Posteroanterior
- **PCA** - Principal Component Analysis
- **PGGAN** - Progressive Growing Generative Adversarial Network
- **PPG** - Photoplethysmogram
- **PSNR** - Peak Signal-to-Noise Ratio

- **Q**
- **QUANTUM** - (Referenced as a consortium setting data quality standards)

R

- **REALM** - (Project name - specific acronym meaning not explicitly provided)
- **REPL** - Read-Eval-Print Loop
- **RIANA** - (Dashboard name - specific acronym meaning not provided)
- **RMSE** - Root Mean Squared Error
- **RNN** - Recurrent Neural Network
- **RN** - Required Number (for disease generation)
- **ROUGE** - Recall-Oriented Understudy for Gisting Evaluation
- **RWD** - Real-World Data

S

- **SAPBERT** - Self-alignment Pretraining for BERT
- **SDQF** - Standard Data Quality Framework
- **SOTA** - State-Of-The-Art
- **SPADE** - Spatially-Adaptive (normalization)
- **SSIM** - Structural Similarity Index Measure

T

- **tabDDPM** - Tabular Denoising Diffusion Probabilistic Model
- **TN** - True Negative
- **TRTR** - Train on Real, Test on Real
- **TRTS** - Train on Real, Test on Synthetic
- **TSTR** - Train on Synthetic, Test on Real
- **t-SNE** - t-distributed Stochastic Neighbor Embedding
- **TWED** - Time Warp Edit Distance

U

- **UM** - (Partner organization - University of Maastricht implied)
- **UQI** - Universal Quality Index

V

- **VAE** - Variational Autoencoder
- **VITO** - (Partner organization name)

- **VPHI** - (Partner organization name)
- **VQ** - Vector Quantization
- **VQ-GAN** - Vector-Quantized Generative Adversarial Network
- **VQ-VAE** - Vector-Quantized Variational Autoencoder

W

- **WGAN** - Wasserstein Generative Adversarial Network
- **WGAN-GP** - Wasserstein Generative Adversarial Network with Gradient Penalty

Executive Summary

Deliverable D4.6 "Multimodal Synthetic Data Generator" reports on the outcomes of Task 4.4, carried out between months 5 and 35 of the REALM project. The task was designed to achieve the main objectives: to generate synthetic multimodal medical data and complete patient profiles, to improve accessibility of clinical data with privacy guaranteed.

The work presented in this deliverable has demonstrated that these objectives have been fulfilled. Synthetic data has been successfully generated across all modalities required for the REALM use cases, including time-series, CT, chest X-ray, and electronic health records. Privacy-preserving mechanisms such as differential privacy have been integrated into the synthetic data generation framework, ensuring that the data can be shared and reused without compromising confidentiality.

A distinctive achievement of Task 4.4 is the development of a comprehensive evaluation framework that assesses the quality of synthetic data across multiple dimensions: fidelity, utility, diversity, fairness, and privacy.

A central purpose of the synthetic data framework in Task 4.4 is to enable more accurate subgroup evaluation. By conditionally generating samples for under-represented demographic and clinical strata, the framework alleviates the small-sample problem that typically undermines fairness assessment. The resulting evaluation suites yield narrower confidence intervals, greater statistical power, and more stable estimates of performance across intersections (e.g., age × sex × ethnicity) while preserving privacy through differential privacy and strict non-memorization safeguards. In practice, synthetic data supplements hold-out real data, providing calibrated, modality-specific test sets that allow equitable and evidence-based validation of models across time-series, CT, chest X-ray, and EHR modalities

Beyond methodological advances, Task 4.4 has delivered tangible outputs in the form of open-source code, evaluation benchmarks, and synthetic datasets integrated into the REALM data repository in full compliance with FAIR principles. These outputs provide a lasting contribution to the consortium and to the wider research community.

The results of Task 4.4 are closely linked to other REALM activities, feeding into the statistical and deep learning approaches of Task 5.1, supporting the data fusion and integration work of Task 5.2, underpinning the RIANA dashboard in Task 5.3, and operating within the federated data repository established in Task 4.3. Integration with the REALM architecture and the AI orchestrator (Tasks 3.1 and 3.2) ensures that synthetic data generation is embedded within the overall system design.

The next steps will focus on the complete integration of synthetic data generation into the REALM production environment. This will occur within **Work Package 6**, where all REALM components will be brought together into the REALM evaluation platform. The platform will provide the means to conduct rigorous, privacy-preserving evaluations of the five REALM use cases, ensuring that the innovations developed in Task 4.4 contribute directly to clinical translation and regulatory impact.

In conclusion, this deliverable demonstrates that multimodal synthetic data generation is feasible, reliable, and impactful. By combining advanced generative methods, formal privacy guarantees, and subgroup evaluation for fairness, Task 4.4 has fulfilled its objectives and positioned synthetic data as a cornerstone of the REALM framework for trustworthy healthcare AI.

1 Introduction

This document constitutes Deliverable D4.6, entitled “Multimodal Synthetic Data Generator”, as specified in the REALM project proposal. It presents the results of the activities carried out in Task 4.4 (T4.4), which was conducted between month 5 and month 35 of the project MAINLY by UM, with contributions from VITO, VPHI, MINDS, EXUS, and TB.
According to the project proposal, Deliverable D4.6 provides a comprehensive generation framework that integrates data generators for various medical data types and includes the design of evaluation metrics and benchmarks to assess the overall performance of the framework. The deliverable consists of two main components:
The deliverable comprises two principal elements: firstly, the open-source code for the generators, evaluation methods, and benchmarks, which are hosted externally and will be referenced in this document; and secondly, the scientific report (presented here)that details the design and evaluation of multimodal synthetic data generation. Together, these outputs ensure that the deliverable not only advances the technical objectives of Task 4.4 but also contributes lasting, reusable assets to the REALM ecosystem in full compliance with FAIR data principles.

1.1 Deliverable Scope

1.1.1 Determining the Task Scope

The scope of Task 4.4 encompasses the development of methods and frameworks for the generation and validation of synthetic multimodal clinical data, as well as the advancement of digital twin simulations. Importantly, synthetic data generation is a foundational element of the REALM platform, designed specifically to facilitate the evaluation of submitted medical software devices that implement AI models. Within the Federated Cloud-based Data Repository, synthetic data generation enables secure, privacy-preserving, and comprehensive testing environments, particularly in situations where real-world data is limited, restricted, or does not represent the full spectrum of patient diversity.
Strategically, the Data Synthesis component addresses key limitations inherent in traditional evaluation approaches by:

- Augmenting scarce real-world datasets for underrepresented subgroups identified in our studies, thereby improving the robustness and inclusivity of device evaluations.
- Enabling privacy-preserving evaluation through differentially private electronic health record (EHR) generation and the inherent privacy protections of synthetic image and time-series data, ensuring compliance with data protection regulations.
- Supporting comprehensive bias assessment via subgroup-level evaluation across multiple modalities, allowing for the identification and mitigation of potential biases in AI-driven medical device software.

Further, Task 4.4 explores the integration of synthetic data with digital twin models to simulate patient profiles exhibiting a wide range of characteristics, including those not sufficiently represented in available real-world datasets. By addressing modalities such as time series data, computed tomography scans, X-ray images, and electronic health records, Task 4.4 ensures all data requirements for the REALM use-cases are met. The scope additionally includes the design of robust evaluation metrics and benchmarks to assess the fidelity, utility, diversity, fairness, and privacy guarantees of the generated data, ensuring that outputs are technically reliable, clinically relevant, and aligned with regulatory expectations.
In summary, synthetic data generation within the REALM platform is not only pivotal for advancing the technical objectives of Task 4.4, but it also directly underpins the evaluation framework for medical device software, safeguarding patient privacy and enabling rigorous, fair, and representative assessment of AI models in healthcare.

1.1.2 Relation to Other REALM Tasks

Task 4.4 has not been carried out in isolation but is closely interlinked with other REALM activities. In particular, T4.4 is strongly connected to Task 5.1, which addresses post-market evaluation of synthetic data. The outcomes of T4.4 contribute directly to the work of Task 5.1 by providing synthetic datasets needed for evaluation. Furthermore, T4.4 contributes to the evaluation activities undertaken in Work Package 6, where the generated synthetic datasets are assessed in the context of the REALM use-cases. Finally, the synthetic datasets and evaluation benchmarks developed in T4.4 are integrated into the REALM data repository, thereby ensuring compliance with FAIR data principles and enabling reuse within and beyond the consortium.

1.1.3 Relation to REALM use cases

The deliverable also underlines the relevance of synthetic data for the REALM use-cases. These use-cases require data across several modalities, including time series, computed tomography (CT), X-ray imaging, and electronic health records (EHRs). By covering all these modalities, the work in T4.4 ensures that the consortium can develop and validate solutions in line with clinical needs while mitigating the risks associated with the use of sensitive patient data.

1.1.4 Direct Fulfilment of REALM Promises

The activities conducted in Task 4.4 directly fulfil the promises made in the REALM project proposal. The framework developed under this task enables the **generation of synthetic multimodal medical data** across several domains, including imaging modalities such as CT and chest X-ray, time-series data such as intensive care unit monitoring, and structured electronic health records. By incorporating differential privacy techniques and leveraging the inherent privacy-preserving properties of generative models, the task has demonstrated how **personal medical data can be made accessible while safeguarding patient confidentiality**. Moreover, the conditional generation capabilities of the framework allow the creation of targeted patient populations with specific demographic or clinical characteristics without privacy risks.

The work also demonstrates strong alignment with the REALM use-cases. Synthetic CT scan generation directly supports **Use Case 1** on lung cancer evaluation, where conditional adversarial networks for medical image data generation were promised. Time-series generation contributes to **Use Case 3** on glucose control and **Use Case 4** on COPD inpatient risk stratification, using advanced statistical solutions to extrapolate and augment real-world distributions. Similarly, EHR generation underpins Use Case 4, which focuses on COPD EHR-based models, and **Use Case 5**, which addresses COPD patient-reported outcomes.

From a technical perspective, Task 4.4 has delivered several specific contributions that directly address commitments made in the proposal. The implementation of privacy-preserving synthetic structured data generation using DP-CGANs fulfills the promise of **enabling users to balance privacy preservation with data utility**. Advanced medical image generation methods for CT and chest X-ray data realize the commitment to **develop adversarial deep architectures for medical image synthesis with controllable conditions**. **Rare disease scenarios** are addressed through ontology-guided generation, which enables the training and testing of medical algorithms in cases where data is extremely scarce. In addition, a comprehensive evaluation framework has been developed to assess fidelity, utility, diversity, and fairness of synthetic data, meeting the promised requirement of **rigorous evaluation benchmarks**.

Finally, the achievements of Task 4.4 extend beyond the state of the art. The Enhanced TimeAutoDiff model surpasses standard approaches by delivering substantial improvements in utility through innovative distribution alignment penalties. Cross-domain pretraining between chest X-ray and CT demonstrates novel transfer learning capabilities. Formal privacy guarantees with quantifiable trade-offs between privacy and utility exceed conventional claims in the synthetic data domain. Moreover,

subgroup evaluation techniques enable fairness assessments across demographic intersections that were previously challenging. Collectively, these achievements illustrate how the framework developed in Task 4.4 fulfils the REALM promise of enabling advanced data analytics and decision-making on large-scale medical data, while ensuring privacy preservation and supporting realistic clinical scenarios for medical device software evaluation.

1.1.5 Key contributions

This deliverable synthesizes comprehensive research efforts within the REALM project, focusing on the development and application of multimodal synthetic data generators. Our work makes several key contributions to the field:

1. A **comprehensive systematic review** of generative AI models for synthetic data across multiple medical modalities.
2. **Enhanced time-series generation** through novel diffusion model architectures with distribution alignment penalties
3. **Demographically-conditioned medical image synthesis** across chest X-rays and CT scans
4. **Privacy-preserving EHR generation** with formal differential privacy guarantees
5. **Cross-modal evaluation frameworks** for assessing synthetic data utility in fairness evaluation
6. **Practical integration** within the REALM platform architecture for real-world deployment

1.2 Challenges in traditional source of regulatory evidence

The healthcare industry stands at the threshold of an AI revolution. Machine learning models are increasingly deployed across medical domains, from diagnostic imaging and clinical decision support to personalized treatment recommendations and drug discovery. This transformation promises unprecedented improvements in patient outcomes, clinical efficiency, and healthcare accessibility. However, realizing this potential requires AI systems that are not only accurate but also fair, robust, and trustworthy across the full spectrum of patient populations.

As outlined in the REALM project proposal, REALM regulatory assessment framework champions the use of multiple data sources to evaluate the performance of medical device software. While clinical trial data, particularly from randomized controlled trials, are highly sought after for regulatory assessment, they are also a major source of challenge. These data are limited to a predefined, restricted population, making them time-consuming, expensive, and complex to acquire. This limitation introduces residual uncertainty for regulators and health technology assessors, who must assess safety and efficacy based on observations from a narrow population. Ultimately, this creates evidence gaps that hinder a comprehensive assessment of a medical product's effectiveness in a broader patient population.
The constraints of clinical trial data are often mitigated by retrospective **real-world data** (RWD) sourced from registries and routine clinical practice. These data sources accumulate regulatory evidence over a longer time period and can cover a wider population than limited clinical trials. While there is increasing interest in using RWD for regulatory assessments, these data also have limitations, including inconsistent quality, scarcity, and a lack of representativeness across diverse patient populations, reflecting existing biases in healthcare delivery. These issues prevent reliable assessment of model fairness and may mask algorithmic biases that could perpetuate or exacerbate health disparities. Furthermore, the use of RWD is often hindered by **data privacy concerns** .

1.3 REALM's solution for evidence gaps

1.3.1 Access to real-world data pools

In the above regulatory context, **REALM** addresses evidence gaps in conventional clinical trial data by complementing them with data from public registries and real-world data from public data banks like the **UK Biobank**. T4.3 – Information sharing with European medical data pools, and deliverable D4.5

(M9), outline the various such initiatives for sourcing real-world data. As a result of these data-sharing agreements concluded with collaborating partners, REALM has access to a **federated cloud-based data repository** (D4.3) that facilitates the use of real-world data for regulatory assessment of AI models in medical device software. Access to both internal and external data pools within REALM is monitored and managed through an active **Data Management Plan** (DMP) – see D4.2 M6, D4.3 M30.

1.3.2 Assessing quality of data

To gauge potential inconsistencies in the quality of real-world data from independent public repositories, REALM has established a **Standard Data Quality Framework** (SDQF). This framework outlines 24 dimensions, consistent with the quality standards set by the QUANTUM consortium, that is setting up pathway to meet **European Health Data Space** regulations. To operationalize these data quality metrics, REALM has implemented qualitative and quantitative metrics to benchmark data quality for both AI model development and evaluation within medical device software. Deliverable D4.4 on Standard Data Quality Framework, maps the data quality with regard to the AI model evaluation metrics like technical performance, data governance, transparency and accountability, as well as fairness.

Data quality Score

The technical implementation of REALM's **Data Quality Framework** uses quantitative metrics to inform stakeholders about potential data quality gaps and inconsistencies. The resulting **"Data Quality Score"** helps to assess the impact of these issues on the downstream use of data for evaluating the performance of AI models during the regulatory assessment of medical device software. For instance, a low score in the **population representativity dimension** could indicate a significant class imbalance. Similarly, a low **metadata granularity score** would reflect a lack of detailed metadata for the data sources.

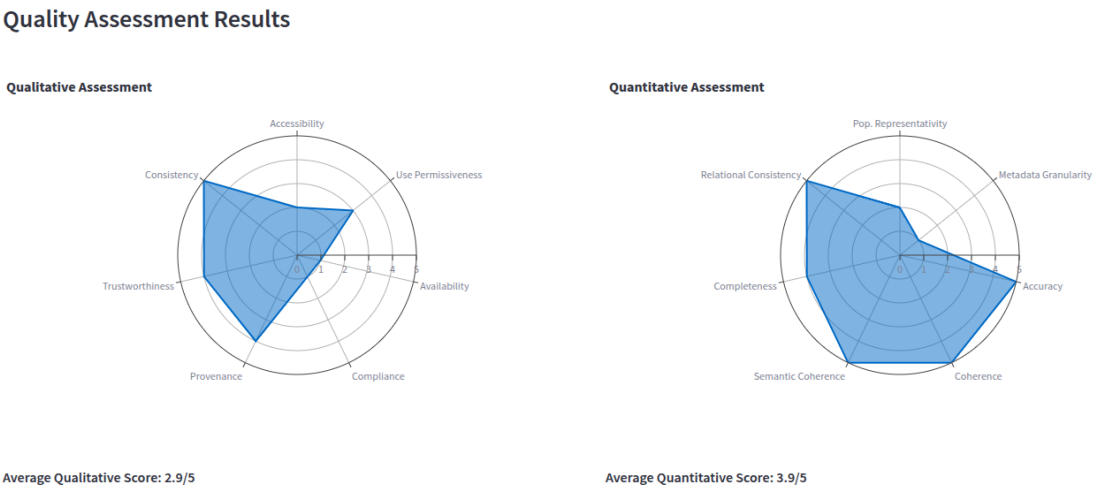

Figure 1. Radar diagrams illustrative the data quality scores, as graded by REALM's data quality framework.

Deliverable D4.4 provides further details on the dimensions of data quality and the technical implementation of REALM's data quality scores.

While assessing the quality of data used in regulatory assessment is helpful, mitigating its shortcomings is paramount. **REALM** addresses constraints such as data inconsistency, scarcity, and a lack of diverse population representation by using **synthetic data** while still preserving privacy.

1.3.3 Synthetic data

Briefly, synthetic data generators are computational models that create **in silico** (computer-generated) data that mimic real patient data. This synthetic data is expected to maintain the physiological statistical, functional properties as real patient cohorts but lacks personal information, as well as constraints associated with data quality and representation.

Since synthetic data are generated from multiple patients and are designed to represent conditions associated with complex diseases, they can replicate observations found in real clinical subjects. By varying the "input parameters" of the synthetic data generator, we can emulate real observations. The wide variety of synthetic data generator models developed in **Task 4.4** allow the REALM framework to meet the extensive needs of different data types (time-series, tabular, images) and the desired sample size required for regulatory assessment. Thus, synthetic data serves as a substitute for real-world data, overcoming challenges related to data sharing, cost, and delays. By creating artificial datasets that statistically mimic real-world data while containing no direct patient identifiers, synthetic data offers several advantages:

- **Enhanced Data Sharing:** Facilitates the sharing of sensitive medical information for research and development without compromising patient privacy.
- **Data Augmentation:** Supplements limited real datasets, enabling the training of more robust and generalizable AI models.
- **Bias Mitigation:** Allows for the creation of balanced datasets by oversampling underrepresented subgroups, thereby addressing algorithmic bias.
- **Robust Model Evaluation:** Provides a controlled environment for rigorously evaluating AI model performance, especially in scenarios where real test data is scarce or sensitive.

1.3.4 Digital twin simulations

While synthetic data broadly can serve as a substitute for real-world data, **Digital twin simulations are high quality imputations** that build on the synthetic data, to test different clinical conditions and treatment options through computer simulations. More importantly, digital twins can predict outcomes that are difficult or impossible to measure with physical experiments and can also forecast how outcomes will evolve over time in response to changing external conditions.

Digital twin simulations are **virtual experiments** conducted using **digital twins**, which are virtual representations of a physical object or system. When **emulated synthetic data** is filtered for clinical applicability and validated for a specific context of use, these synthetic data cohorts can represent the realistic characteristics of a target patient population. These "clinically plausible virtual patient cohorts" can then be used to perform high-fidelity computer simulations with digital twins. This process allows for the comparison of different therapy combinations, the prediction of treatment progression and outcomes, and even patient stratification for real clinical trials.

Within the **REALM** framework, **digital twin simulations** are explored to support the regulatory assessment of AI models. By accessing a library of digital twin models that replicate organs, pathologies, or treatment options, REALM framework could conduct virtual simulations that provide high-quality clinical context and insights to evaluate a model's safety and performance.

1.3.5 REALM's pathway for leveraging digital twin simulations

REALM's vision for leveraging digital twin simulations is to connect the REALM platform with a European repository of digital twins in health, foreseen within the **Ecosystem of Digitial Twins in Healthcare (EDITH) -Coordinate & Support Action[3] and related EC initiative- Virtual Human Twin[4] infrastructure**. **Annex 1** provides an overview of various existing digital twin resources identified by the EDITH-CSA. These are potential candidates for a future virtual human twin infrastructure, which could also be used by platforms like REALM to run digital twin simulations.

While the EDITH-CSA project has provided an extensive roadmap for a multi-scale, multi-organ, multi-pathology digital twin ecosystem, the tender for building the public infrastructure to host these digital twin computer models is still being evaluated[5]. REALM consortium members VITO and VPHi, who are

[3] EDITH-CSA – An ecosystem for digital twins in Health care - EDITH project s a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement No. 101083771 - https://www.edith-csa.eu/

[4] Virtual Human Twins initiative for health and care - https://digital-strategy.ec.europa.eu/en/policies/virtual-human-twins

[5] Call for Tender: Virtual Humans Twins Platform Project – opening 25/08/2025- closing – 28/07/2025 - LINK

also core members of EDITH-CSA and Virtual Human Twin Community of practice, continue to closely follow the above developments on public infrastructure on digital twins.

The technical development of digital twins for high-fidelity simulations was intended to be realised through possible integration with EDITH infrastructure. The originally envisioned public infrastructure of digital twins likely remains unavailable within the REALM's life-time. Therefore, we here briefly present a curated list of digital twin models reproduced from the EDITH-CSA initiative (EDITH-CSA - Deliverable – Proof of concept on available resources[6]). This includes descriptions and workflows for digital twins intended for personalized simulations across **four use cases**: **Intensive Care Unit (ICU)**, **Osteoporosis**, and **Cancer (see Annex A1 – Table A1)**. Additionally, a brief list of potential digital twins and/or computational models curated by EDITH initiative is summarised in Annex 1 - Table A2.

2 Foundations: A Systematic Review of Generative AI in Medicine

Having established how our work directly fulfills REALM's commitments and supports the project's five use cases, we now turn to examining the current landscape of generative AI approaches. This systematic review provides the foundational understanding necessary to appreciate the innovations presented in subsequent sections and demonstrates our comprehensive approach to advancing beyond the current state-of-the-art.

Before presenting our novel contributions, it is essential to understand the current landscape of generative AI in healthcare. This systematic review provides the foundation for appreciating both the opportunities and limitations that motivated our research directions.

A comprehensive systematic review was conducted to provide a holistic understanding of generative models (GANs, VAEs, Diffusion Models, and LLMs) used to synthesize various medical data types. Unlike previous narrowly focused reviews, this study encompassed a broad array of medical data modalities, including imaging (dermoscopic, mammographic, ultrasound, CT, MRI, and X-ray), text, time-series, and tabular data (EHR).

[6] EDITH-CSA – An ecosystem for digital twins in health care. Deliverable D5.1– PoC repository with available resources (EDITH – GA No. 101083771)

2.1 Overview of Generative Models

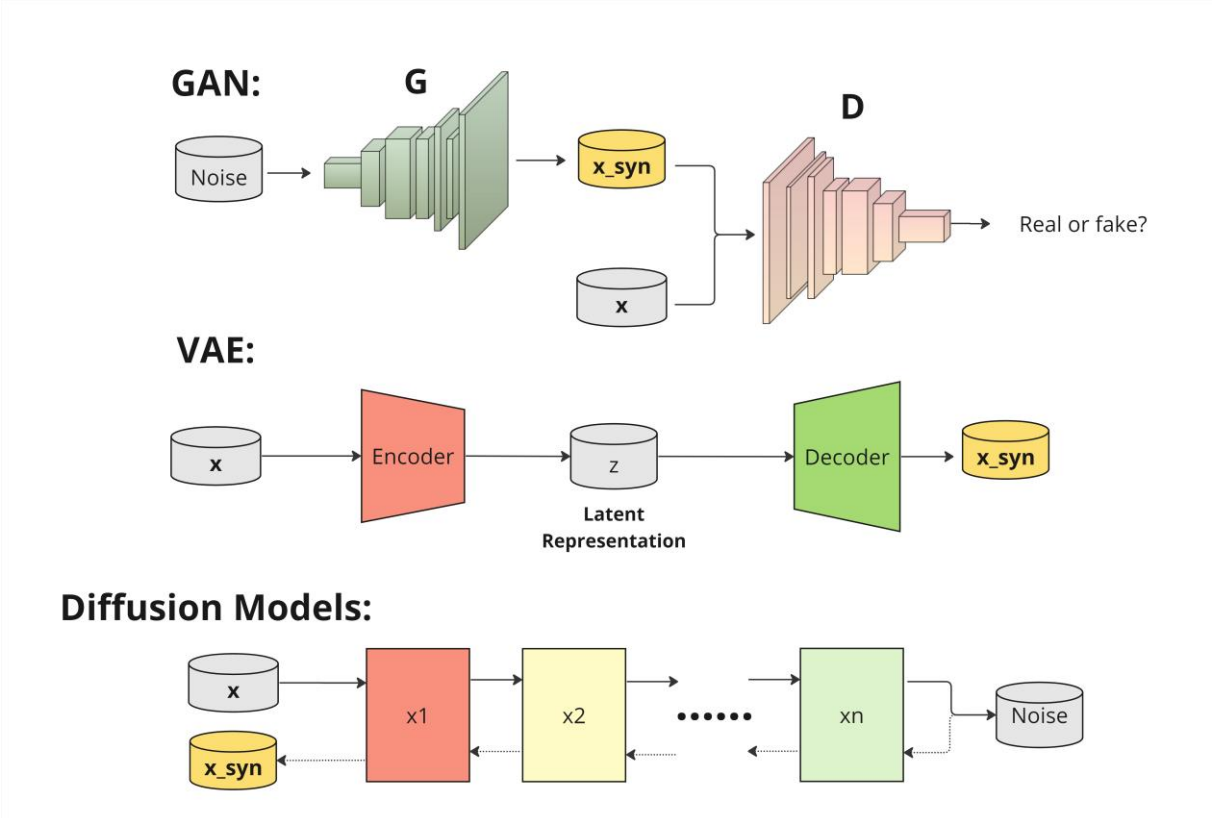

Figure 2-Generative Models. GANs follow adversarial training to implicitly model the real distribution. VAEs employs variational inference techniques to approximate real distribution, while diffusion models gradually add noise to the input and reverse it back.

The review encompassed a broad spectrum of generative AI architectures, including:

- **Generative Adversarial Networks (GANs):** Models that learn to generate new data instances that resemble the training data by pitting two neural networks (a generator and a discriminator) against each other.
- **Variational Autoencoders (VAEs):** Probabilistic graphical models that learn a latent representation of the input data and can generate new data by sampling from this latent space. VAEs consist of an encoder and a decoder network.
- **Diffusion Models (DMs):** A class of generative models that learn to reverse a gradual diffusion process, transforming noise into data. These models have shown remarkable success in generating high-fidelity images.
- **Large Language Models (LLMs):** While primarily designed for text, LLMs are increasingly being adapted for medical text generation and even for guiding image synthesis.

See Figure 2 for the basic working mechanisms of the different generative models.

2.1.1 Differences in generative models working mechanism per datatype

- **Handling discrete variables:** GANs, initially introduced for generating realistic continuous images, face challenges when applied to discrete data. Similarly, many diffusion models are Gaussian processes, operating in continuous spaces and not the discrete space. To address these issues, various techniques have been introduced. (Variational) autoencoders are commonly employed before GANs and diffusion models to condense high-dimensional and heterogeneous features into latent representations. Notably, some diffusion models like Tabular Denoising Diffusion Probabilistic Models (tabDDPM) adopt a mixed sequence diffusion approach, employing Gaussian diffusion for continuous variables and multinomial diffusion for

discrete variables. Others treat discrete variables similar to real-valued sequences but with further post-processing of the model output, transforming continuous variables into discrete.

- **Handling time aspect**: Notably, conventional vanilla GANs and diffusion models lack the ability to generate time-series data, making them unsuitable for directly creating time-dependent EHR data or signals. As a remedy, specialized sequential Deep Learning (DL) models like Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-term Memory (LSTMs) are employed within the architectures of the GANs or diffusion models.

2.1.2 GANs

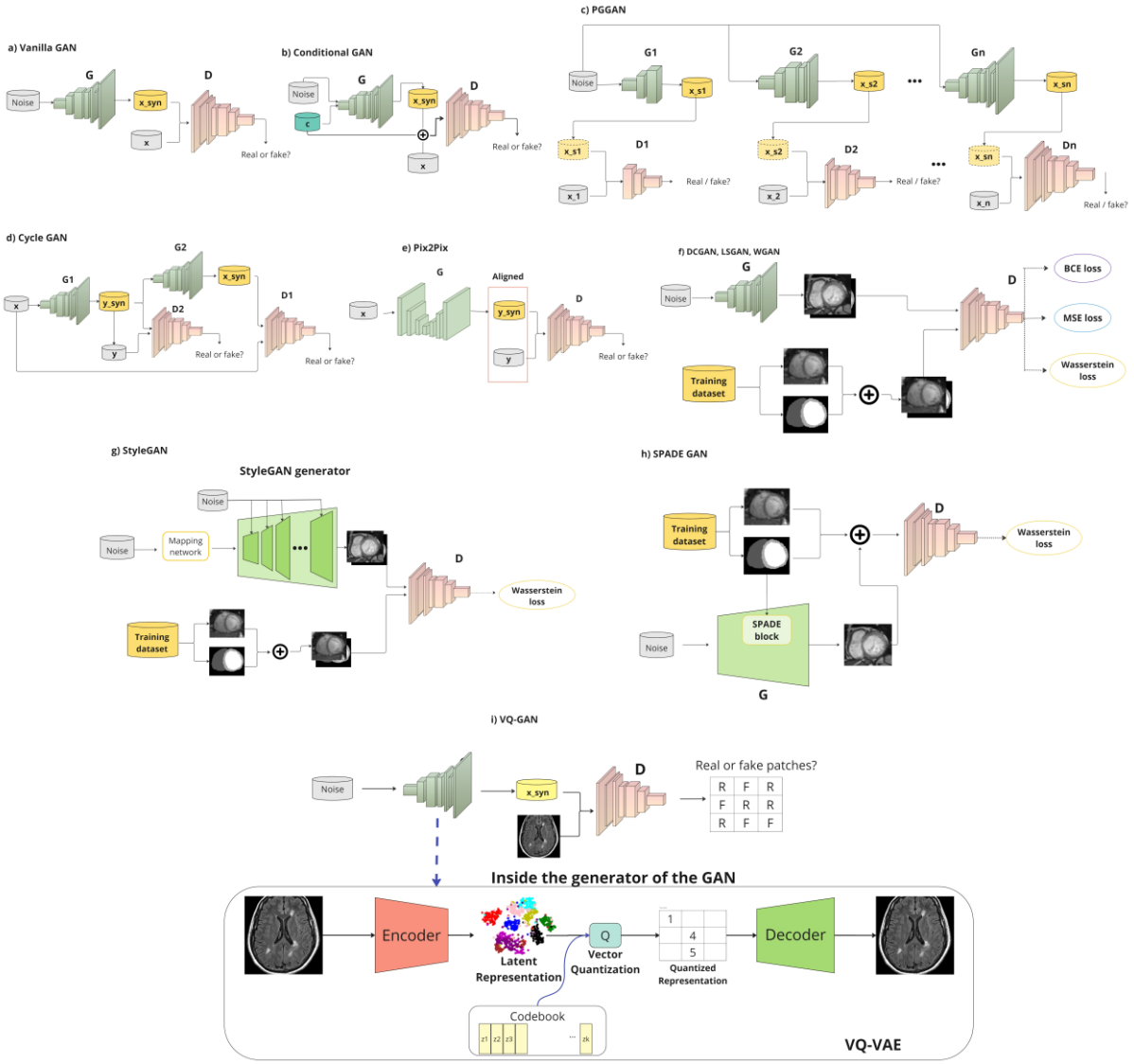

*Figure 3- Comprehensive overview of various GAN architectures, each tailored for specific applications and improvements in synthetic data generation. The **Vanilla GAN** (a) illustrates the basic GAN structure with a generator creating samples from noise and a discriminator evaluating them. The **Conditional GAN** (b) integrates conditional variables to guide the generation process, enabling the production of more targeted outputs. The **PGGAN** (c) employs a series of generator and discriminator pairs (G1, D1 to Gn, Dn) that progressively increase the resolution and detail of the generated samples. **CycleGAN** (d) features a dual-generator system for effective style transfer, where each generator learns to translate images between two distinct domains, and the outputs are cycled back as inputs to maintain consistency. The {Pix2Pix} (e) uses aligned pairs of images for precise image-to-image translation tasks, converting segmented images to realistic photos using a trained generator. The Generator G is usually an encoder-decoder net or U-Net with skip connection. **DCGAN**, **LSGAN**, and **WGAN** (f), use different loss functions like BCE, MSE, and Wasserstein to refine the discriminator's accuracy. **StyleGAN** (g) features a sophisticated generator with a mapping network that inputs noise, refines the generation through style blocks, and uses a discriminator with Wasserstein loss. **SPADEGAN** (h) utilizes spatially-adaptive normalization to modulate the synthesis based on the input segmentation map, with a discriminator employing Wasserstein loss. **VQ-GAN** (i) merges GAN and VAE elements by encoding inputs into a latent space, quantizing them for efficient representation, and then reconstructing the output to improve image*

quality. x and x_syn as well as y and y_syn denote input and synthesized data pairs, while c represents the conditional variable. Similarly, x_1 and x_s1, x_2 and x_s2, x_n and x_sn represent real and synthesized data pairs.

Adversarial approaches, such as GANs, involve a generator and discriminator, trained to outperform each other, hence the term "adversarial". The generator network adapts its distribution and generates a new, reliable distribution, while the discriminator network learns to differentiate between real and augmented data.

The work in Goodfellow et al., 2014 pioneered GANs for generating realistic images, particularly dealing with continuous data, often referred to as the **vanilla GAN** (Figure 3a). However, such models have been extended to different domains, such as audio and EEG signal generation (Donahue et al., 2018; Raoof & Gupta, 2021). (Radford et al., 2015) proposed the **Deep Convolutional GAN (DCGAN)**, that combines the original GAN with convolutional neural networks for better and more stable training. (Arjovsky et al., 2017) proposed the **Wasserstein GAN (WGAN)**, while (Mao et al., 2017) proposed the **Least Squares GAN (LSGAN)**, to address challenges like mode collapse and vanishing gradients typically encountered in GANs (see Figure 3f).**WGAN** replaces binary classification with the Wasserstein distance. The training of WGAN was enhanced with the introduction of **WGAN-GP** in (Gulrajani et al., 2017), which integrates a gradient penalty technique. **Progressive Growing GAN (PGGAN)** (Karras et al., 2017) is an extension of GAN training, ensuring stability for generators producing large, high-quality images. Traditional GANs struggle with stability when working with larger sizes, attempting to balance both structure and fine details. This challenge worsens as resolutions increase, often required for medical image generation, leading to training failures, while memory constraints on Graphics Processing Units (GPUs) further necessitate reducing batch size, introducing instability. As illustrated in Figure 3c, PGGAN addresses these issues by incrementally increasing model size during training, starting small and gradually adding layers until achieving the desired image size.

GANs enable the generation of diverse image, video, or audio data from random input sampled from a normal distribution(Alqahtani et al., 2021). However, vanilla, unconditional GANs lack control over the generated samples' appearance or class. To address this, **Conditional Generative Adversarial Networks (cGANs)** are introduced, which allow for conditional generation by incorporating semantic input or a condition (c), like image class, into the generation process (Mirza & Osindero, 2014), as shown in Figure 3b. Both the generator and discriminator in cGANs are conditioned on auxiliary information, enabling the model to learn complex mappings from diverse contextual inputs to corresponding outputs.

Conditional GANs have popularized the use of image conditioning in image-to-image translation tasks. For such tasks, models like **Pix2Pix** (Isola et al., 2018) (see Figure 3e) and its high-resolution counterpart (T.-C. Wang et al., 2018) rely on supervised training with paired datasets. **Pix2PixHD** addresses some limitations of Pix2Pix, specifically catering to high-resolution image translation. **CycleGAN** (Zhu et al., 2020) offers an alternative by excelling in unsupervised image translation, eliminating the need for paired datasets. CycleGAN utilizes cycle consistency, allowing generated images from one generator to serve as input for the other, with the output matching the original image, enabling consistency in both directions, as shown in Figure 3d. This capability proves valuable in scenarios where obtaining paired training data is challenging. While CycleGAN shines in unsupervised learning, it is not optimal for high-resolution image-to-image translation. Consequently, paired approaches are often preferred for medical data generation (T. Wang et al., 2021)

Other notable GAN-based approaches frequently employed for medical data generation, include **StyleGAN** (Karras et al., 2019), **SPADE GAN** (Park et al., 2019), and **VQ-GAN** (Esser et al., 2021). Style Generative Adversarial Network (StyleGAN) (Karras et al., 2019), depicted in Figure 3g, represents a significant extension to the GAN architecture. It introduces changes like a mapping network for latent space, allowing control over style at various points in the generator model, and the incorporation of noise for variation. The StyleGAN generator and discriminator models are trained using the PGGAN

method. This model not only produces high-quality, high-resolution images but also provides control over style at different detail levels through manipulation of style vectors and noise. A cutting-edge image translation model, akin to Pix2Pix, is SPADE GAN, which introduces spatially-adaptive (SPADE) normalization. Unlike previous models, which employed segmentation maps only at the input layer, SPADE GAN ensures the segmentation map is inputted across all intermediate layers, preserving its information throughout the model's depth. (see Figure 3h).

The VAE introduces a framework for compressing data into a lower-dimensional space (Kingma & Welling, 2022). The Vector-Quantized Variational Autoencoder (**VQ-VAE**) model enhances representation by mapping inputs to a continuous space and then discretizing these into tokens, facilitated by an encoder, decoder, and a learnable codebook (Oord et al., 2017). The Vector-Quantized Generative Adversarial Network (VQ-GAN) fuses VAE and GAN models with vector quantization (VQ) to create data, including images, from noise, and employs VQ to elevate the quality by turning outputs into discrete values. This technique not only improves the definition and edges of generated data but also provides examples like images with more precise and clearer contours compared to outputs from conventional GANs, benefiting from the combined strengths of GANs and VQ's detailed representation (Esser et al., 2021).

2.1.3 Diffusion Models

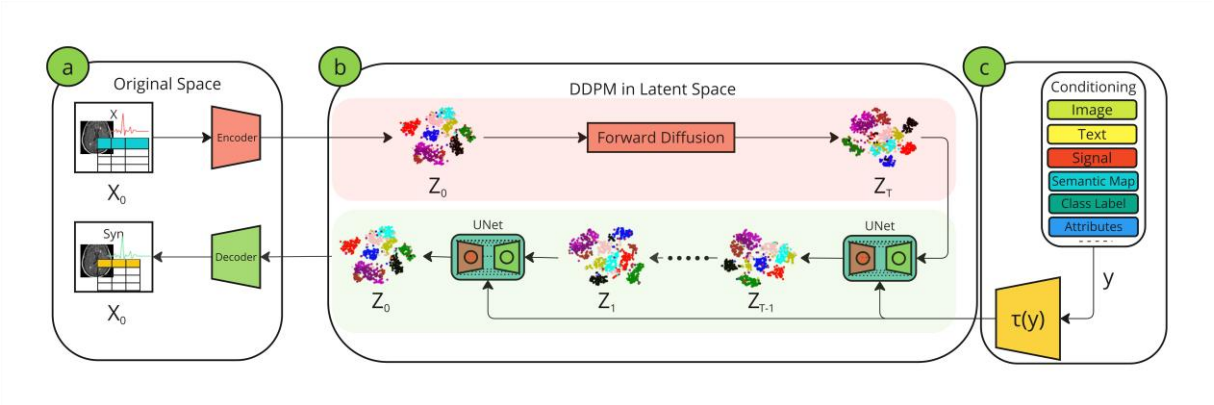

Figure 4-Diffusion Models. ***a,b,c*** *constitute the components of the Stable diffusion pipeline.* ***a****: (Variational)AutoEncoder component for compressing the high-dimensional inputs X_0 into the lower-dimensional latent representations Z_0 and mapping them back to the pixel space. X_0 can be an image, a signal, or a table with heterogeneous features.* ***b****: DDPM process consisting of a forward diffusion process and a backward one. Z_0 is destructed by adding noise iteratively. Z_0 is then reconstructed from Z_T iteratively using a neural network. Within the pipeline of Latent Diffusion Model, the DDPM process is performed in the latent space. Originally in the DDPM, the diffusion is performed in the original space.* ***c****: Optionally, an embedded conditioning tau(y) can be added, where y could be other images, radiological reports, semantic maps, class labels... In SD, the embedding function tau is based on CLIP*

The original concept of the diffusion model, introduced by Diffusion Probabilistic Models (DPMs) (Sohl-Dickstein et al., 2015), draws inspiration from non-equilibrium statistical physics. The key idea is to systematically and iteratively destroy structure in a data distribution through a forward diffusion process. Subsequently, the reverse diffusion process is learned and applied to restore the structure in the data.

DDPM

The initial practical implementation of the diffusion model in the context of images was presented by (Ho et al., 2020), introducing Denoising Diffusion Probabilistic Models (DDPM). This approach destroys data by iteratively adding Gaussian noise to the image according to the Markov chain. To learn the reverse process, a deep neural network is needed to recover the data. Figure 4b illustrates the working mechanism of the DDPM. The current best practice for image diffusion models is to use U-Net (Ronneberger et al., 2015) architectures for denoising. However, these architectures are tailored for image generation tasks, and may not be a viable option for other data types, especially when the

training data is limited. Moreover, U-Net architectures struggle to retain temporal dynamic information and lack flexibility in handling varying input sequence lengths, thus proving inadequate for managing sequential information. In addressing this limitation, some studies (Tian et al., 2023) resort to Bidirectional Recurrent Neural Networks (BRNNs) implemented with either LSTM or GRU units. A significant drawback of diffusion models is that they operate sequentially on the entire image during both training and inference. Consequently, they demand substantial computational resources and time, making both the training to be slow as well generating an image after training. As a result, (Song et al., 2022) speed up image generation by redefining the diffusion process as a non-Markovian one, which allows for skipping steps in the denoising process, without requiring all past states to be visited before the current state.

Latent Diffusion Models- The Stable Diffusion Pipeline

To address computational inefficiencies of DDPMs, Latent Diffusion Model (LDM) (Rombach et al., 2022) was introduced initially for images, which uses encoders to compress images from their original size in the pixel space into a smaller representation in the latent space, as illustrated in Figure 4a. Therefore, in latent diffusion, the diffusion process is performed on the latent representations rather than the original images (Figure 4b), allowing LDMs to model long-range dependencies within the data and enabling training on limited computational resources while retaining their quality and flexibility. {Stable diffusion}, a foundation model built based on LDM, introduces text conditioning to the model for additional control over the generation process and consists of three main components: (i) the VAE for compression (Figure 4a), (ii) a U-net based diffusion process in the latent space (Figure 4b), and (iii) a conditioning mechanism that embeds a prompt describing the image using a Contrastive Language-Image Pre-training (CLIP) text encoder (Radford et al., 2021) (Figure 4c). CLIP creates a numeric representation (embedding vector) of the prompt, mapping both the text and images into the same representational space, enabling comparison and similarity quantification. In addition to CLIP, other text encoders such as pre-trained T5X model (Roberts et al., 2022), medBERT encoder (Rasmy et al., 2020), a Byte Pair Encoding (BPE) tokenizer(Sennrich et al., 2016), and Self-alignment pretraining for BERT (SAPBERT) (Liu et al., 2021) are commonly used in different publications to enable text conditioning of the models.

The Stable Diffusion model is widely utilized in recent publications, where the pre-trained foundation model is fine-tuned on various medical modalities without the need to initiate training from scratch. Moreover, the idea of latent diffusion models, which involves performing the diffusion process in the latent space, has been extended to both the EHR and signals data types. In these extensions, a low-dimensional representation of the high-dimensional features, including discrete ones, is learned using an auto encoder before the diffusion process.

2.2 Applications Across Medical Modalities

Our systematic review revealed distinct patterns in how generative models are applied across medical data types.

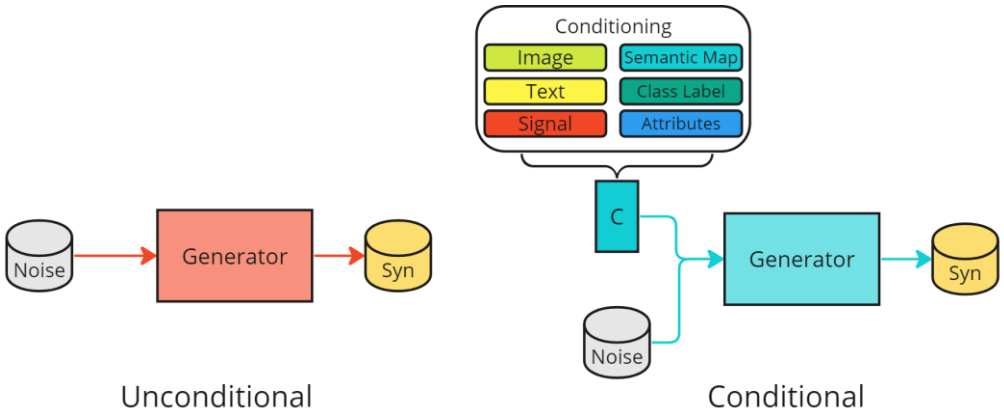

Figure 5- Unconditional (left) vs conditional generation (right)

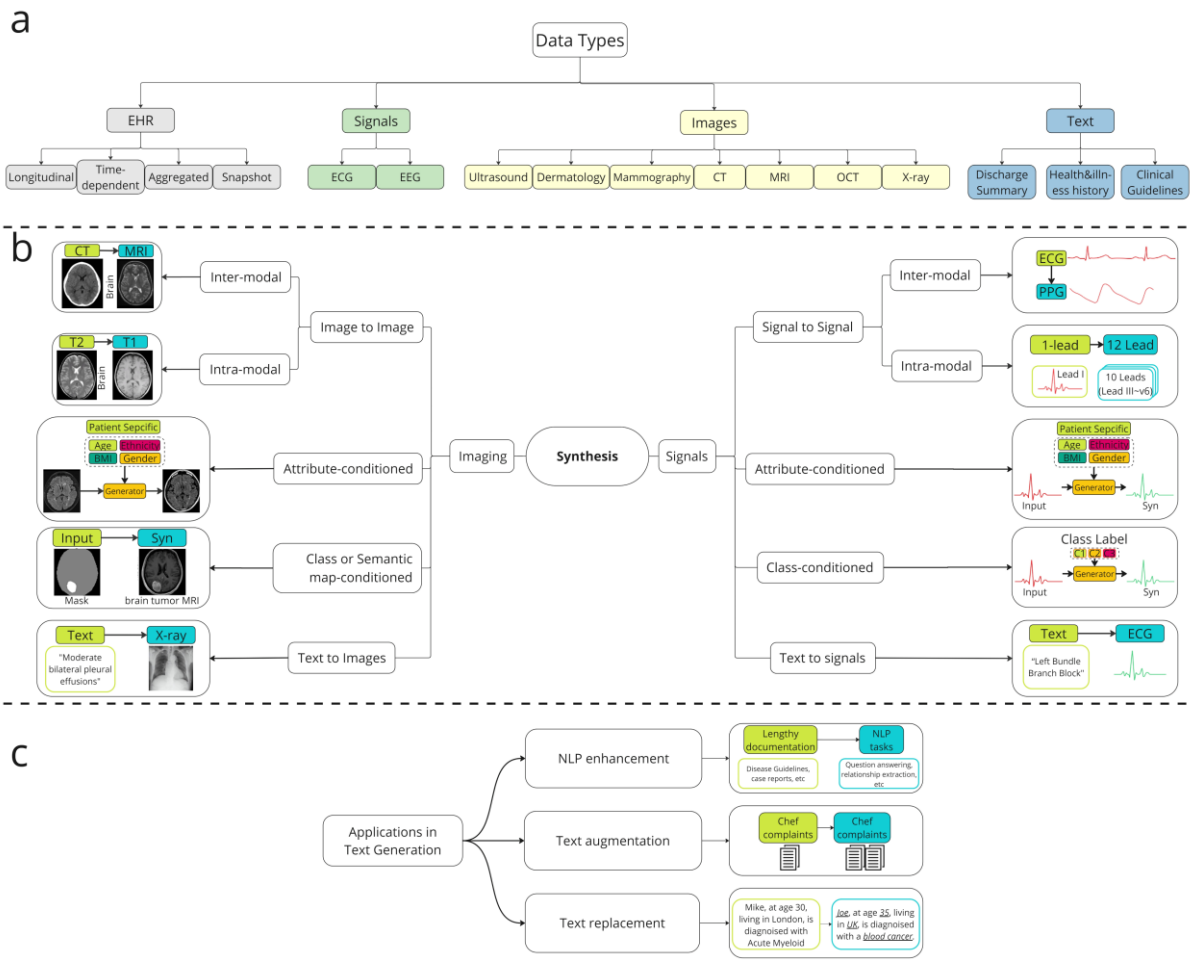

Figure 6- (a) The different data types and modalities covered in the survey. The different EHR formats also represent the synthesis applications of EHR. (b) Overview of synthesis applications of the imaging and signals data type. The left side focuses on imaging data, while the right side focuses on signals. An example of inter-modal synthesis involves the transformation of a brain CT scan to an MRI in imaging or an ECG signal into a synthetic PPG signal in the signals data type, which helps in scenarios where certain imaging or signal modalities might be unavailable. Intra-modal synthesis involves the conversion of a brain MRI T2 sequence into a T1 sequence or transforming a single lead into a 10 lead synthetic ECG signal. Attribute-conditioned synthesis shows a patient-specific brain MRI or an ECG signal being generated that matches attributes like age, BMI, ethnicity, and gender. Class or semantic map-conditioned synthesis shows a synthetic brain MRI with tumor being generated using a binary mask of a brain tumor or a synthetic ECG signal being generated based on class labels, such as C1 and C2. This can be useful for generating labeled datasets for training. The figure also illustrates text-guided synthesis, such as generating a synthetic chest X-ray based on a textual description like "moderate bilateral pleural effusions" or a synthetic ECG signal based on textual descriptions like "Left Bundle Branch Block". (c) Synthesis applications of the text data type

We introduce specific definitions for the terms "data types" and "modalities". We classify EHR, images, text, and signals as distinct data types. Within each data type, we define various modalities. For example, within medical images, modalities include CT, MRI, and others. Additionally, within each modality, there exist different modalities; for example, within MRI, modalities such as T1 and T2 weights are distinguished.

Generative models are a powerful tool in machine learning and can be broadly categorized into two types: unconditional and conditional. Unconditional models take a random variable as input, allowing for the application of unconditional synthesis. On the other hand, conditional generative models introduce an additional layer of control by incorporating external information or context during the generation process that serve as additional guidance for the model. These can include images, text, semantic maps, class labels, attributes, and signals, as demonstrated in Figure 5. This added control allows conditional models to be used in a variety of synthesis applications specific to each data type shown in Figure 6.

2.2.1 EHR Applications

For the EHR data type, the synthesis depends on the EHR format, which includes:

- **Longitudinal EHR:** Consisting of medical codes from various patient visits.
- **Aggregated EHR:** Longitudinal visits of the patients condensed into a single row.
- **Time-dependent EHR:** Time-series readings from one patient's visit.
- **Snapshot EHR:** A single snapshot of a patient's EHR focusing on selected attributes.

2.2.2 Physiological Signals Applications

Regarding the physiological signals and imaging data types, prominent applications include

- **Inter-modal translation:** Converting data from one modality to another, such as translating CT scans to MR images in the realm of imaging, or converting ECG readings to PPG signals in the signals data type.
- **Intra-modal translation:** Translating data within the same modality, like converting different MRI contrasts (e.g., T1-weighted images to T2-weighted images) or converting single-lead ECG inputs into a complete 12-lead set.
- **Class or semantic map-based synthesis:** Generating data conditioned on a certain class label or a specific segmentation mask.
- **Attribute-based synthesis**: Creating data conditioned to correspond with specific characteristics and patient demographics, such as age, sex, and Body Mass Index (BMI). These approaches enable a higher level of personalization in the synthetic data, potentially leading to more accurate representations of patient-specific medical data.
- **Text-based synthesis:** Demonstrating an innovative approach for integrating descriptive text into synthetic medical data generation by conditioning on clinical text reports. This could involve text-to-image, text-to-signal, or text-to-table synthesis.

2.2.3 Medical Text Applications

For the medical text synthesis, we categorize the applications based on their generation purposes into

- **Natural Language Processing (NLP) enhancement:** The generated synthetic text data is used to improve the performance of NLP tasks. Typical NLP tasks in medical text generation include name entity recognition, information/concept extraction, relation extraction, semantic similarity, summarization, and question answering. For example, the Name Entity Recognition (NER) performance on real clinical notes is limited by the insufficient amount and completeness of the dataset. Adding generated synthetic medical text can improve the NER task in this scenario.

- **Text augmentation:** The generated text is used to augment existing text such as generating discharge summaries, patients reports, clinical notes, and case reports of certain diseases, especially when real-world clinical notes are limited and sensitive to share or use for research.
- **Text de-identification**: By replacing personal identifiable and sensitive information about individuals, this application preserves patient privacy in clinical text. These applications focus on particular attributes (such as names, addresses, and diagnoses) in the text.

2.3 Overview of Evaluation Metrics

The evaluation of synthetic data is a multifaceted task, encompassing several dimensions that collectively determine the quality and usability of the synthetic data. We categorize the evaluation as utility, fidelity, diversity, qualitative assessment, clinical validation, and privacy. Each dimension presents unique challenges and considerations, and their comprehensive evaluation is crucial to ensure the effectiveness of synthetic data.

2.3.1 Evaluation of Utility

Evaluation of utility asses the use of the synthetic data for specific tasks, such as providing additional data for improving a downstream medical AI model. Several settings exist in which the synthetic data can be used for a downstream task. The synthetic data can be used to fully train the downstream model (replacing the need to access real data), or to augment real training data. Moreover, the synthetic data can be used for validating or testing the models. These settings can be employed in different downstream applications such as supervised classification, prediction, supervised regressions, image segmentation, even reinforcement learning, with the specific metrics of each application.

2.3.2 Evaluation of Privacy

Evaluation of privacy is a critical factor that ensures the synthetic dataset cannot be exploited to reveal sensitive information about individuals in the original dataset. When it comes to EHR data, privacy evaluation typically includes testing the synthetic data for the following risks:

- Membership inference risk: The risk that an attacker can determine whether a specific data record was used to train a machine learning model.
- Attribute inference risk: The risk that an attacker can infer private attributes, such as age, sex, ethnicity, income, of individuals using the available data or a model's predictions.
- Nearest neighbour adversarial accuracy risk: This risk is evaluated using the Nearest Neighbour Adversarial Accuracy (NNAA) metric (Yale et al., 2020). It assesses privacy by measuring how closely the nearest neighbour in the synthetic dataset matches any real individual. High similarity increases the chance of re-identification.

Researchers usually apply Differential Privacy (DP) (Dwork, 2008) mechanisms to mitigate privacy risks. DP is a rigorous and mathematically grounded framework designed to protect the privacy of individuals in the dataset. By introducing randomness into data queries or the data generation process, differential privacy ensures that the synthetic data doesn't compromise the privacy of any individual in the original dataset. It's often considered the gold standard for privacy protection in data analysis.

2.3.3 Evaluation of Fidelity

Evaluation of fidelity assesses how well the synthetic data reflects the real-world characteristics and features of the original dataset, both at the individual sample level and across the entire data distribution. High fidelity indicates that the synthetic data closely resembles the original data in terms of appearance and statistical properties. Notably, fidelity metrics differ across different data types due to their inherent structural and inherent differences. Figure 7 summarizes fidelity metrics for different types of medical data. In most of the cases, the fidelity metrics used are originally designed for different types of data like natural images, not taking into account the characteristics and statistics of typical medical image modalities. For instance, FID (Fréchet Inception Distance) and IS (Inception Score) are

widely used to evaluate the similarity between synthetic and real images, but they may not accurately reflect the subtle variations and noise patterns inherent in medical data, as these metrics are underpinned by a feature extractor pre-trained on the ImageNet dataset, which is more relevant for natural images.
Sure, I've transcribed the table into your document. Here it is:

Modality	Purpose of evaluation	Fidelity Metric
EHR	Dimension wise distributional similarity	Bernoulli Success probability Chi-square test Kullback-Leibler divergence (KLD) Kolmogorov-Smirnov Test Pearson test
	Joint Distribution Similarity	Jensen-Shannon Divergence Maximum-Mean Discrepancy (MMD) Wasserstein distance
	Inter-dimensional Relationship Similarity	Propensity score Pearson pairwise correlation Pairwise correlation difference Dimension-wise prediction
	Latent Distribution similarity	Log-cluster metric Latent space representation
	specific-language model	Perplexity
	specific-time-series	Autocorrelation function Patient trajectories
Imaging Data	Pixel-wise similarity (image based)	(Root/Normalized) Mean squared error (MSE, RMSE, NRMSE) Structural similarity index (SSIM) Multi Scale SSIM (MS SSIM) Peak signal-to-noise ratio (PSNR) Universal Quality Index (UQI) Contrast Noise Ratio (CNR)
	Feature-wise similarity (distribution based)	Fréchet inception distance (FID) Kernel Inception Distance (KID) Inception score (IS) Maximum Mean Discrepancy (MMD) Learned Perceptual Image Patch Similarity (LPIPS) Feature Similarity Index (FSIM) feature distribution similarity (FDS)
	Image/text Alignment	Bilingual Evaluation Understudy (BLEU) Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Contrastive Language-Image Pre-training score (CLIP)
Time-series Data	Temporal sequences similarity	Dynamic Time Warping (DTW) Time Warp Edit Distance (TWED)
	Temporal Correlation	Autocorrelation Cross-correlation (CC)
Medical Text	Similarity between machine and human translation	Bilingual Evaluation Understudy (BLEU) Cosine similarities
	Similarity between real and generated text	Jaccard Similarity ROUGE-N recall G2 - test

Figure 7. Fidelity Metrics

2.3.4 Evaluation of Diversity

Evaluation of diversity measures the extent to which synthetic data captures the full variability of the real data. Following (Alaa et al., 2022), we consider diversity and fidelity as distinct yet complementary metrics in evaluating synthetic data. This ensures that the generative models are not only evaluated for reproducing average or common scenarios but also generating the broad spectrum of real-world variations in the original RW dataset.
Figure 8 summarizes diversity metrics for different types of medical data.

Modality	Diversity Metrics
EHR	Category coverage Generated Disease types Required sample number to generate all diseases (RN) Normalized Distance Medical concept abundance
Signals	Signals Diversity of Samples Score
Imaging	MS-SSIM Nearest SSIM difference Visualization(Diversity) Diversity score (DS) LPIPS Hammind Distance Euclindean Distance

Figure 8. Diversity Metrics

2.3.5 Clinical Evaluation

Clinical evaluation involves assessing the synthetic data's correct representation of anatomical details consistent with real-world clinical scenarios, its ability to replicate key clinical features and characteristics, and its adherence to established medical guidelines and practices. These anatomical and biological details may be overlooked by fidelity metrics, and thus a clinically relevant evaluation is pivotal in introducing synthetic images into the medical field. Clinical validation also aim to verify that synthetic data can effectively support medical decision-making, patient care, and clinical outcomes, ultimately enhancing its utility and trustworthiness in healthcare applications.

2.3.6 Qualitative Evaluation

Qualitative evaluation including human assessment, also known as a visual Turing Test, which involves medical experts reviewing the synthetic data to try to distinguish it from the real data. In some cases, qualitative evaluation only involves data visualization using techniques such as t-distributed Stochastic Neighbor Embedding (t-sne) and Principal Component Analysis (PCA), to compare the distribution of real and synthetic data visually.

2.4 Key Insights and Identified Gaps

Our broad, multi-modal review has uncovered several unique findings that would not have been evident through single-modality surveys. By analysing generative models across various types of medical data, we reveal opportunities and challenges that span modalities while identifying shared gaps and transferable solutions:

- **Core techniques vs. modality-specific needs:** The core techniques of generative models are consistent across modalities, but data preparation is modality-specific.
- **Need for personalized, context-aware synthesis:** Shared gaps, such as the limited integration of patient context, clinical knowledge, and task-specific requirements, highlight the need for more personalized, context-aware synthesis methods.

- **Cross-modality inspiration:** Techniques successful in one domain, such as textual conditioning in imaging, can inspire advancements in others, such as EHR or signal synthesis, revealing untapped potential for cross-modality innovation.
- **Underexplored applications:** Synthetic data are underutilized beyond augmentation; it can support validation, edge case testing, and explainability, improving model robustness, and fostering clinician trust.
- **Evaluation challenges:** The lack of robust evaluation frameworks is a universal challenge, emphasizing the need for cross-modality benchmarks, diversity metrics, and large-scale clinical validation.
- **Modality-specific challenges:** Different modalities have unique requirements: imaging demands attention to physics-based properties, EHR synthesis must preserve temporal dependencies and handle mixed data types, and signals require attention to morphology and noise.
- **Potential of multimodal data synthesis**: Multimodal synthetic datasets (e.g., imaging with EHR or text) offer promising opportunities for comprehensive diagnostic and predictive applications.
- **Benefits of text-guided conditioning across modalities**: Text-guided conditioning is emerging as a versatile technique applicable to both imaging and non-imaging data, bridging structured and unstructured data in generative processes.
- **Scaling synthesis for clinical use:** Under-explored areas, such as physiological signals and multimodal data, require improved synthesis techniques, computational scalability, and more attention to real-world clinical validation across all modalities.

Based on the insights collected from the surveyed papers, we propose the following recommendations to be taken into account for future research:

- **Generative approaches should be tailored to the medical data and downstream task or their intended use:** As specified earlier, generative approaches should be more tailored to the specifics of the medical data and also to the specific task at hand, as there are different requirements for each task. For instance, requirements for segmentation/classification are substantially different compared to registration or denoising tasks. Furthermore, medical imaging data, such as CTs and MRIs, are acquired through physics-based processes, like X-ray attenuation and k-space signal processing, which shape their structure and texture. Recent works leverage k-space properties and diffusion models to enhance image quality, preserve texture, and accelerate sampling. We recommend further exploration of conditional models and task-specific priors to improve generative approaches across modalities.

- **Synthetic data should be utilized beyond data augmentation to train and evaluate medical AI software:** Synthetic data should not only serve as a tool for data augmentation and training downstream models but also for validation and testing purposes. Initiatives like DARWiN EU and REALM EU demonstrate its use in regulatory decision-making, ensuring privacy while evaluating model performance under diverse scenarios. Moreover, it can be used for the development of explainability mechanisms to improve not only the confidence of the proposed analysis models, but also the confidence of clinicians in utilizing automated methods in clinical practice.
- **Prioritizing fairness and inclusivity in medical data generation:** Due to data scarcity, there is a need to focus on more personalized synthesis tailored to specific subject characteristics and patient demographics, prioritizing fairness and inclusivity in medical data generation. Recent approaches offer promising solutions by integrating sensitive attributes like race and gender to ensure balanced representation and generate more inclusive data. These strategies emphasize the importance of fairness-aware synthesis methods that address demographic disparities and foster equity in clinical applications.

- **Adapting language models to medical tasks should be further approached with more care:** A thorough analysis is needed on the interpretability capabilities of diffusion models when it comes to the understanding and translation of radiological reports to images. Notably, emphasizes the importance of aligning LLMs with domain-specific knowledge. Building on this, we recommend further exploring interpretability techniques, such as visualization of decision pathways, to bridge the gap between radiological text reports and medical images.

- **A call for benchmarking and comparative studies to promote openness and collaboration**: The use of established and well defined benchmarks and open-source datasets would not only accelerate progress in the field but also provide valuable insights into the strengths and weaknesses of various techniques, guiding researchers towards the optimal tools for specific clinical applications.

- **A need for more comprehensive evaluation:** There is a need for holistic assessment covering more specific fidelity metrics, proper clinical validation and human assessment, and well-defined diversity metrics. Similar efforts have been presented, which lays out a framework that can be followed to establish a consistent set of metrics and evaluation practices of generative data. Furthermore, there is a need for well established privacy evaluation practices, especially for data types where there is a significant gap. This is crucial for ensuring the responsible deployment of synthetic data in medical contexts.

These gaps directly motivated the research directions presented in the following sections, where we address fairness, privacy, and clinical utility as primary design objectives.

3 Enhanced Time-Series Generation: Supporting REALM Use Case 3&4

Having established the current state of generative AI in medicine and identified key limitations, we now turn to our first major contribution: developing enhanced synthetic data generation specifically designed to enable reliable subgroup-level evaluation in time-series settings.

3.1 Context and Problem

Our time-series synthetic data generation directly addresses two critical REALM use cases: **Use Case 3 (Glucose control in intensive care units - STAR)** and **Use Case 4 (COPD inpatient risk stratification using time-series model).** This work fulfills the project's promise to provide "advanced statistical solutions that extrapolate and augment RWD distributions enabled by REALM's proposed renewable certification dataset for (a) new numerical data generation (e.g., blood glucose, insulin levels)."

Both use cases require continuous monitoring data crucial for patient outcome prediction yet face identical challenges REALM committed to address: limited data for underrepresented patient subgroups and the need for privacy-preserving evaluation frameworks. Traditional evaluation approaches often rely on small, unbalanced test sets that provide unreliable performance estimates for minority populations, potentially masking critical disparities in model performance. Our approach directly implements the promised capability to enable "training and testing medical algorithms in realistic scenarios" while addressing "scenarios with little-to-no available samples."

REALM Task 4.4 Implementation: This work specifically delivers on the commitment to "capture information from the interaction network as well as the omics datasets, and fuse them to generate omics level synthetic data with reliable predictive signals" by developing Enhanced TimeAutoDiff that captures complex temporal relationships in ICU monitoring data while preserving demographic-specific patterns essential for fair model evaluation across both glucose control and COPD scenarios.

3.2 Enhanced Time AutoDiff: Beyond State-of-the-art Achievement

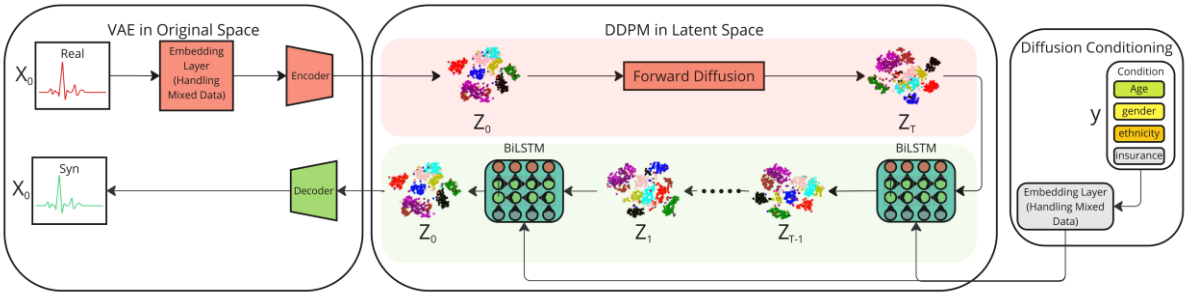

Figure 9. Overview of the Enhanced TimeAutoDiff architecture.

Our Enhanced TimeAutoDiff directly fulfills REALM's promise to advance **"Beyond SoTA"** by developing novel distribution alignment mechanisms that address a fundamental limitation in existing synthetic data approaches: the gap between training utility (TSTR- train on synthetic, test on real) and evaluation utility (TRTS- train on real, test on syntehtic).

Building on prior diffusion and VAE-based generators (TimeDiff, HealthGen, TimeAutoDiff), we introduced **Enhanced Time AutoDiff**, a novel framework that augments the latent diffusion objective with distribution-alignment penalties. This innovative approach aims to ensure that the synthetic data's underlying distribution closely matches that of the real data, leading to more reliable evaluations.

Technical Innovation Delivering Project Promises: Building upon the project's commitment to "adversarial deep architectures," we present Enhanced TimeAutoDiff, an improved version of the TimeAutoDiff generator.To encourage a well-behaved latent space, we augment each base loss with two additional penalties:

- **Maximum Mean Discrepancy (MMD) Loss:**

$$\ell_{MMD} = \frac{1}{B^2}\sum_{\{i,j\}} k(x_i, x_j) + \frac{1}{B^2}\sum_{\{i,j\}} k(y_i, y_j) - \frac{2}{B^2}\sum_{\{i,j\}} k(x_i, y_j)$$

where $k(\cdot,\cdot)$ is a radial basis function kernel. This ensures distributional alignment between real and synthetic representations, directly addressing the project's requirement for "statistical similarity between real and synthetic data"

- **Consistency Regularization:**

$$\ell_{Consistency} = MSE\big(f(x), f(x + \delta)\big), \qquad \delta \sim N(0, \sigma^{2I})$$

This enforces smooth model behavior under small input perturbations, improving robustness of generated samples.

- The overall enhanced losses become:

$$L_{Auto}^{Enhanced} = L_{Auto} + \lambda_{Auto}^{MMD}\ell_{MMD} + \lambda_{Auto}^{consistency}\ell_{consistency}$$

$$L_{\text{Diffusion}}^{Enhanced} = L_{\text{Diffusion}} + \lambda_{\text{Diffusion}}^{MMD}\ell_{MMD} + \lambda_{\text{Diffusion}}^{consistency}\ell_{consistency}$$

The model trained with these enhanced objectives is our Enhanced TimeAutoDiff. This architectural advancement directly implements the promised "generation framework" that "combines data generators of various medical data types" while ensuring the "privacy preservation level and data utility level" balance specified in **Task 4.4.**

3.2.1 Baseline Models

In addition to the TimeAutoDiff baseline, we compare to two models: TimeDiff and HealthGen. TimeDiff (Tian et al., 2023) is a diffusion probabilistic model that also uses a bideriectional reccurent neural network (BiRNN) architecture for realistic privacy-preserving EHR time series generation. To simultaneously generate both continuous and discrete-valued time series, TimeDiff introduce a mixed diffusion approach that combines multinomial and Gaussian diffusion for EHR time series generation.

The HealthGen (Bing et al., 2022) model consists of a dynamical VAE-based architecture that allows for the generation of feature time series with informative missing values, conditioned on high level static variables and binary labels.

3.3 REALM-Compliant Evaluation Framework

Our experimental design fulfils multiple REALM deliverable requirements simultaneously:

1. **Use Case 3&4 Validation**: Using MIMIC-III and eICU databases with focus on relevant prediction tasks (24-hour mortality and length-of-stay), directly supporting the use cases system evaluation needs.
2. **Privacy-Preserving Evaluation**: Our framework enables the goal to "improve personal medical data accessibility with privacy guaranteed" by demonstrating reliable model evaluation without requiring access to sensitive real patient subpopulations.
3. **Fairness Assessment Implementation**: Our rigorous subgroup evaluation framework directly delivers the promised "evaluation metrics and benchmarks to assess the quality of generated synthetic multi-model data" by systematically comparing performance across 32 demographic intersections (4 age × 2 sex × 4 ethnicity groups).

3.3.1 Datasets and Tasks

We use the publicly available **MIMIC-III** (A. Johnson et al., 2021) and **eICU** (Pollard et al., 2018) ICU EHRs, which consist of deidentified ICU data linked by patient ID and timestamp. Following established cohort-selection criteria (Van De Water et al., 2023), we extract each patient's first 24 hours of hourly measurements, along with static demographics (age, sex, ethnicity).
We frame two popular downstream tasks on this 24-hour window:

- **ICU Mortality Prediction**: 24-hour mortality risk assessment
- **Binary Length of Stay**: Predicting stays > 3 days

We consider 8 hourly features for the prediction: four vital signs (heart rate, respiratory rate, oxygen saturation, and mean blood pressure), plus four binary indicators for whether each value is missing. Missingness itself is retained via a mask, and missing values are forward-filled. The missingness masks are used as input for the models.
After preprocessing,

- **eICU** yields 113 380 patients (mortality = 5.51 %, LOS > 3 d = 42.3 %)
- **MIMIC** yields 38 129 patients (mortality = 8.13 %, LOS > 3 d = 48.7 %).

For each dataset, we stratify by demographic and outcome, then split into 45 % for training the generative model, 45 % for training downstream classifiers, and 10 % for final hold-out evaluation. This ensures sufficient data for stable generative model training (45%), adequate downstream model training (45%) and validation(10%).

3.3.2 Evaluation Metrics

Our goal is twofold: (1) to improve global evaluation utility (TRTS) so that real-trained models evaluated on synthetic data match performance on true held-out real data, and (2) to enable accurate subgroup-level evaluation by generating large synthetic cohorts conditioned on demographic labels.

Let

$$D_{Real} = \{(x_i, y_i, c_i)\}_{\{i=1\}}^{N}$$

be the real ICU dataset of size N, with time-series x_i, outcome y_i, demographics x_i. We split D_{Real} into disjoint training R^{train} (Size N_{train}) , holdout real $R^{holdout}$, and holdout validation $R^{holdout-val}$ subsets.

Train a generative model G on R^{train} and let S be the generated data of size N_{train} (possibly conditioned on y and c during training). Since R^{train} both trains the generative model and serves as the reference for synthetic comparisons, we will henceforth omit the superscript and denote it simply as R.

Utility for training

Using identical architectures and hyper-parameters, we train two downstream models:

1. $M_{Real} = train(R)$
2. $M_{Syn} = train(S)$

Both models are evaluated on hold-out real data $R^{holdout}$:

$$TRTR_{train} = Evaluate(M_{real}, R^{holdout})$$
$$TSTR_{train} = Evaluate(M_{Syn}, R^{holdout})$$
$$\Delta_{TSTR} = |TRTR_{train} - TSTR_{train}|$$

$TRTR_{train}$ and $TSTR_{train}$ refer to "Train on Real, Test on Real" and "Train on Synthetic, Test on Real," respectively, and are used to evaluate the training utility of synthetic data.

This evaluation is repeated 5 times, each time on a differently trained downstream model.

Utility for Evaluation

We train a downstream predictive model (mortality classifier or LOS classifier) on real holdout dataset $M_{Real} = train(R^{holdout})$. Then, we evaluate on both synthetic and real datasets

$$TRTR_{Evaluate} = Evaluate(M_{real}, R)$$
$$TRTS_{Evaluate} = Evaluate(M_{real}, S)$$
$$\Delta_{TRTS} = |TRTR_{Evaluate} - TRTS_{Evaluate}|$$

$TRTR_{Evaluate}$ and $TRTS_{Evaluate}$ refer to "Train on Real, Test on Real" and "Train on Real, Test on Synthetic," respectively, and are used to evaluate the evaluation utility of synthetic data.

This evaluation is repeated 5 times, each time on a differently trained downstream model.

Utility for Subgroup level Evaluation

For demographic subgroup g, let $R_g \subset R_{train}$ be real samples with subgroup label g. We split R_g (repeated 5 times) into:

- $R_{g,small}$: 20% for simulating small real test sets
- $R_{g,large}$: 80% as ground truth for stable evaluation

We generate synthetic cohort S_g with $|S_g| = |R_{g,large}|$ and compute:

$$\varepsilon_g^{naive} = |AUROC(M_{real}, R_{g,large}) - AUROC(M_{real}, R_{g,small})|$$
$$\varepsilon_g^{synth} = |AUROC(M_{real}, R_{g,large}) - AUROC(M_{real}, S_g)|$$

where ε_g^{naive} represents the error when using small real test sets, and ε_g^{synth} represents the error when using large synthetic cohorts, both compared to the ground truth performance on $R_{g,large}$.

Discriminative Fidelity Score

We split the real and synthetic datasets into train and test subsets:

$$R_{train},\ R_{test} = TrainTestSplit(R_S,\ test\_size = 0.2)$$
$$S_{train},\ S_{test} = TrainTestSplit(S_S,\ test\_size = 0.2)$$

Label every example in R_{train} with 0 (real), and every example in S_{train} with 1 (synthetic). Train a discriminator on this balanced dataset containing R_{train} and S_{train}. Let the discriminator assign a score $p(Synthetic)$ to each sample in the combined dataset ($R_{test} \cup S_{test}$).

$$DiscAUC_S = AUC[y \in \{0,1\},\ p(synthetic)]$$

3.3.3 Demographic Subgroups

We define three categorical demographic dimensions: age bracket (<30, 31-50, 51-70, >70), sex (M and F), and ethnicity (White, Black, Asian, Other). This yields $4 \times 2 \times 4 = 32$ intersectional subgroups of age ×sex ×ethnicity.

To establish reliable reference evaluations, we allocate 80% of each demographic bucket as the `ground truth' set $R_{g,large}$. The downstream model is evaluated on this set to obtain stable performance estimates. The remaining 20% ($R_{g,small}$) simulates the small test samples typically available for minority subgroups in real datasets. The 80/20 split is repeated 5 times with different random seeds to ensure statistical robustness.

3.3.4 Downstream Prediction Models

We implement a one-layer Gated Recurrent Unit (GRU) network followed by a fully connected layer with sigmoid output. Training is done using binary cross-entropy loss, an Adam optimizer (lr = 5e–4), a batch size = 64. We fix these architectures across all experiments for consistency. In all evaluations, we train each downstream model for up to 50 epochs, selecting the best checkpoint via validation performance. We used standard hyperparameters and model architectures from the literature (Bing et al., 2022). Our ablation studies demonstrate that downstream model parameters have minimal effect on the relative performance comparisons between synthetic data generators, ensuring that our conclusions are robust to these choices. The focus of our evaluation is on comparing synthetic data quality rather than optimizing downstream model performance.

3.3.5 Synthetic Data Generation

We implement a standardized generation protocol across four different architectures to ensure fair comparison: Enhanced TimeAutoDiff, baseline TimeAutoDiff, HealthGen, and TimeDiff. While Enhanced TimeAutoDiff, baseline TimeAutoDiff, and HealthGen are inherently conditional models, TimeDiff requires special handling for conditional generation.

Each conditional generator is trained on the same 45% training partition (R_{train}) using identical demographic conditioning. The conditioning vector c includes age bracket, gender, ethnicity, and outcome labels. All generators produce synthetic datasets of equal size (|S| = $|R_{train}|$) with identical conditioning distributions as R_{train}. Specifically, the conditioning labels are drawn to replicate exactly the demographic and outcome distributions observed in the real training data, ensuring that synthetic datasets maintain the same population characteristics.

While TimeDiff is not inherently a conditional model, we adapt it for conditional generation by treating demographic and outcome variables as additional input features during training. This approach embeds conditioning information directly into the learned data distribution. To generate data for specific target conditions, we employ a rejection sampling strategy: we continuously sample from the trained model until we obtain samples that match the desired demographic and outcome criteria.

While computationally less efficient than native conditional generation, this approach ensures that TimeDiff produces synthetic datasets with the same demographic distributions as the conditional models.

3.3.6 Experimental Repetition

We generate 5 independent synthetic datasets for each generator and train 5 downstream models per dataset (real or synthetic), yielding 25 evaluation runs per experimental condition. Moreover, the process of splitting the subgroups into small and large groups is repeated 5 times. This accounts for variability in both synthetic data generation, downstream model training and evaluation, and data splits, enabling robust statistical analysis with reliable confidence intervals and significance testing, while balancing between statistical rigor and computational feasibility. All reported metrics include 95% CIs based on the independent runs.

3.3.7 Alignment weights search

We explored nine configurations that systematically vary the four alignment weights

$$\lambda_{AE-MMD}, \lambda_{AE-consist}, \lambda_{diff-MMD}, \lambda_{diff-consist}$$

across three regimes (zero, light = 0.1, moderate = 0.5). In practice, each experiment requires full diffusion training (on eICU or MIMIC), synthetic sample generation, and downstream training/evaluation. Therefore, this choice of weights and number of configurations ensure sufficient coverage of "no regularization," "light touch," and "moderate emphasis" within a timely possible completion. Moreover, this choice help isolate individual effects and probe pairwise combinations avoiding a combinatorial explosion of configurations

3.4 Key Results

We present our results in three parts. First, we examine discriminative fidelity score on eICU and MIMIC for both mortality and LOS tasks, then we present global evaluation utility (TSTR vs TRTS). Finally, we examine subgroup-level evaluation, comparing small real test sets to pooled real "ground truth" and to large synthetic cohorts conditioned on subgroup labels.

3.4.1 Discriminative Fidelity Comparison

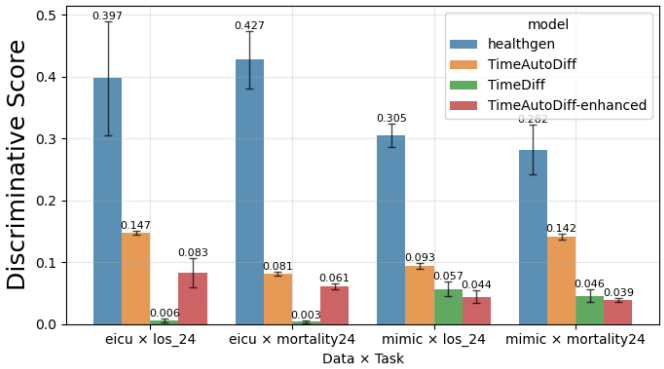

*Figure 10. **Model Comparison - Discriminative Fidelity** This figure displays bar charts comparing the Discriminative Score of HealthGen, TimeAutoDiff, Enhanced TimeAutoDiff and TimeDiff across different tasks and datasets. The lower the scores, the better.*

Figure 3 presents discriminative AUCs for four synthetic data generators: HealthGen (conditional VAE), TimeAutoDiff (autoencoder + diffusion), Enhanced TimeAutoDiff (TimeAutoDiff with extra penalty), and TimeDiff (pure diffusion) on four Data × Task combinations (e.g. eICU LOS_{24}, MIMIC LOS_{24},, eICU $Mortality_{24}$,, MIMIC $Mortality_{24}$).

HealthGen's discriminative scores range from 0.282 to 0.427, meaning a simple classifier can easily distinguish its outputs from real data. TimeAutoDiff improves on this with discriminative AUCs of 0.081–0.147, while TimeDiff performs best, with AUCs as low as 0.003–0.057.
Across every setting, pure diffusion (TimeDiff) is hardest to distinguish from real, followed closely by the Enhanced TimeAutoDiff. The original TimeAutoDiff is an order of magnitude much better than HealthGen, but the added alignment losses yield a substantial drop (25-70%) in the discriminative score compared to the baseline.

3.4.2 Global Utility Comparison

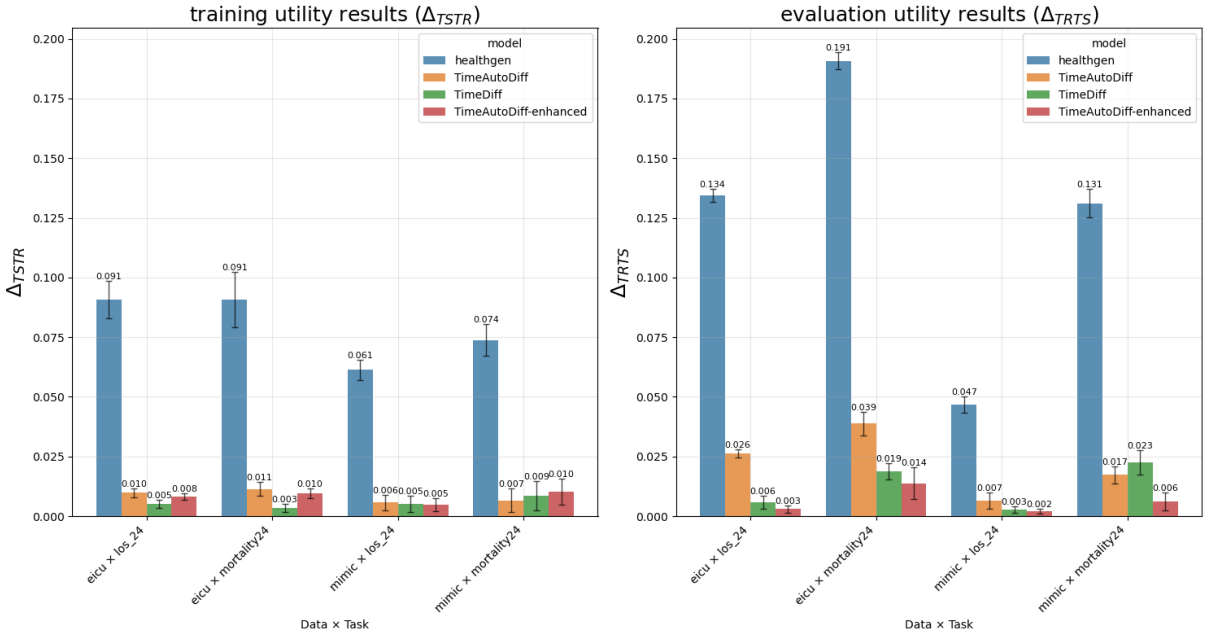

Figure 11.Model Comparison- Global Downstream Utility } This figure displays bar charts comparing the downstream utility for training (Δ_{TSTR})and evaluating (Δ_{TRTS}) of HealthGen,TimeAutoDiff, Enhanced TimeAutoDiff, and TimeDiff across different tasks and datasets. The lower the scores, the better.

We then evaluated the global downstream utility of the four synthetic data generators. As shown in the left-hand panel of Figure 2,HealthGen exhibits large Δ_{TSTR} gaps (≈0.06-0.10), indicating that downstream models trained on HealthGen synthetic data perform substantially worse on real test data than fully real trained baselines. Both TimeAutoDiff and the enhanced variant exhibit similar performance and reduce Δ_{TSTR} to ≈0.006–0.011, and TimeDiff further drives it down to ≈0.003–0.009 across all tasks.
In the right-hand panel, HealthGen again performs worst with Δ_{TRTS} gaps (≈0.13–0.19). TimeAutoDiff cuts those evaluation gaps by roughly 75 % (to ≈0.017–0.039), while TimeDiff drives it down to ≈0.003–0.023. The Enhanced TimeAutoDiff delivers the smallest evaluation gaps of all four: between 0.003 and 0.014 AUROC depending on the task. For instance, on eICU LOS_{24} it achieves just 0.003, and on MIMIC $Mortality_{24}$ 0.006. This shows that our regularized variant of TimeAutoDiff yields the most faithful synthetic hold-outs for downstream evaluation, and when compared to the left-hand panel, this was done at a negligible cost to the generator's ability to produce useful training data.

Comparing the TimeAutoDiff and our enhanced variant of it, we see that the "train on real, test on synthetic" gap shrinks dramatically under the best-enhanced setting. For example, eICU LOS_{24} drops from a gap of 0.026 to 0.003, and MIMIC $Mortality_{24}$falls from 0.017 to 0.006. In each case, the enhanced version (model trained with the optimized weights) reduces Δ_{TRTS} by roughly 70–90%. This means that the synthetic data can give a nearly identical performance as held-out real data for a trained model, all at almost zero cost to the generator's ability to produce synthetic data that train a high-quality downstream model, as the Δ_{TSTR}remains very small (≈0.01) both before and after

enhancement. Therefore, by adding MMD and consistency losses to TimeAutoDiff and by tuning their weights per Data x Task, we significantly reduce the evaluation gap Δ_{TRTS} without sacrificing Δ_{TSTR} or fidelity.

Data	Task	Model	Δ_{TSTR}	Δ_{TRTS}	Relation
eICU	LOS_{24}	TimeAutoDiff	0.01 ± 0.002	0.026 ± 0.002	$\Delta_{TRTS} \approx 3 \times \Delta_{TSTR}$
eICU	$Mortality_{24}$	TimeAutoDiff	0.011 ± 0.003	0.039 ± 0.005	$\Delta_{TRTS} \approx 3.5 \times \Delta_{TSTR}$
eICU	$Mortality_{24}$	TimeDiff	0.003 ± 0.002	0.019 ± 0.003	$\Delta_{TRTS} \approx 6 \times \Delta_{TSTR}$

*Figure 12. **Comparison of Downstream Evaluation and Downstream Training** This table illustrates some examples where the difference between training utility (Δ_{TSTR})and evaluation utility (Δ_{TRTS}) across various tasks and datasets is significantly big.*

Figure 11 and Figure 12 and together show that, even though a downstream model trained on synthetic data (TSTR) almost matches the performance of a model trained on real data, the reverse (training on real and testing on synthetic TRTS) shows a larger discrepancy. For example, looking at the data generated by TimeDiff, specifically for eICU $Mortality_{24}$, Δ_{TSTR} ≈ 0.003 while Δ_{TRTS} is ≈ 0.019, 6 times larger. This problem was substantially mitigated in the Enhanced TimeAutoDiff.

3.4.3 Subgroup-Level Evaluation

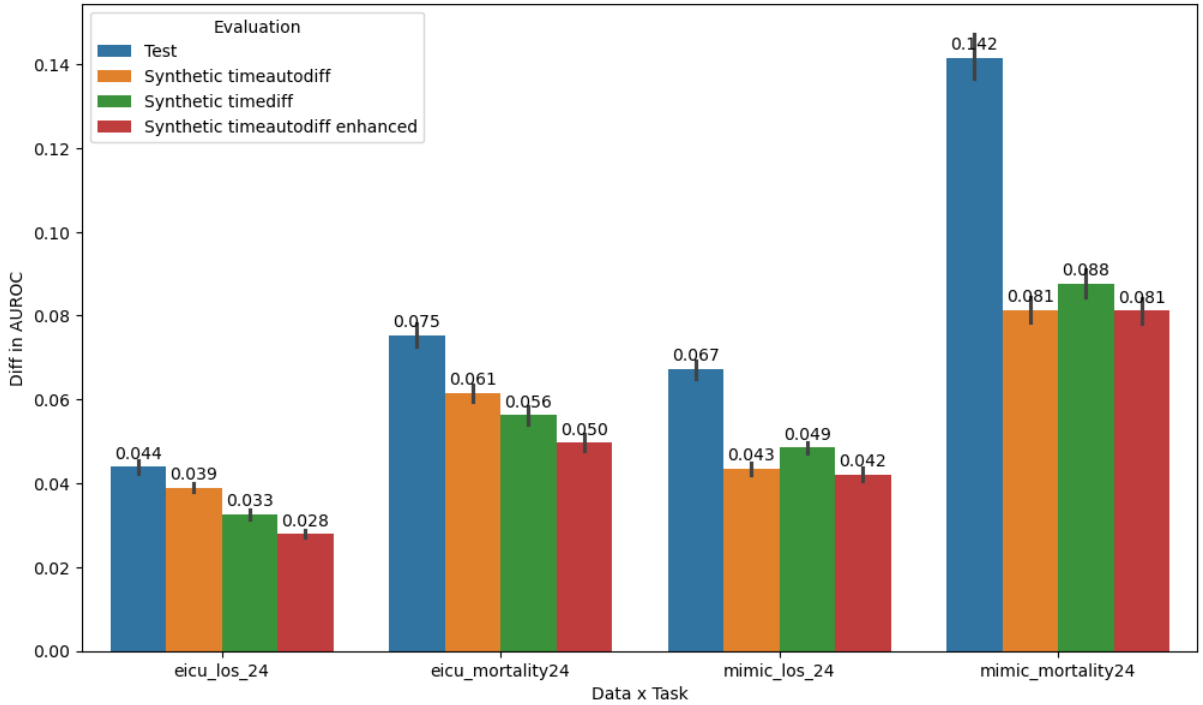

*Figure 13. **Subgroup-Level Evaluation** Mean absolute AUROC error of small-real vs. large-pool ("Test," blue) compared to synthetic cohort vs. large-pool errors using TimeAutoDiff (orange) and TimeDiff(green), and Enhanced TimeAutoDiff (red) averaged over all subgroups for different data and task configurations. The lower the scores, the better.*

For each Data x Task, we compute the absolute AUROC distance between (a) a small real-test subgroup and the large "ground-truth" subgroup (blue bars labeled "Test") versus (b) a synthetic, conditional-on-subgroup cohort and that same ground-truth subgroup (orange, green, and red bars labeled "Synthetic"). In other words, for each subgroup, we compute two errors

$$\varepsilon_g^{naive} = \left|AUROC\left(M_{real}, R_{g,large}\right) - AUROC\left(M_{real}, R_{g,small}\right)\right|$$
$$\varepsilon_g^{synth} = \left|AUROC\left(M_{real}, R_{g,large}\right) - AUROC\left(M_{real}, S_g\right)\right|$$

Data	Task	% of Subgroups Where Synthetic < Test Error		
		TA	TD	TA⁺
eICU	$Mortality_{24}$	48%	60%	76%
eICU	LOS_{24}	44%	52%	72%
mimic	$Mortality_{24}$	68%	68%	84%
mimic	LOS_{24}	72%	56%	76%

Figure 14. Percentage of intersectional subgroups in which the synthetic evaluation error is strictly lower than the real-test error, for each generator: TimeAutoDiff (TA), TimeDiff (TD), and Enhanced TimeAutoDiff TA⁺ across the different configurations.

Then, we average those errors across all subgroups to produce the blue ("Test"), orange ("Synthetic TimeAutoDiff"), green ("Synthetic TimeDiff"), and red ("Enhanced TimeAutoDiff") bars in Figure 13. In Figure 14, we counted, across all subgroups, how often the Synthetic error was significantly smaller than the Test error.
Figure 13 shows that, in every case, the synthetic cohorts cut the error roughly in half compared to the small real-test subsets, and our enhanced generator (Enhanced TimeAutoDiff) yields the lowest errors of all. For example, on eICU LOS_{24} the mean error drops from 0.044 (Test) to 0.039 (TimeAutoDiff), 0.033 (TimeDiff), and 0.028 (Enhanced TimeAutoDiff); on MIMIC $Mortality_{24}$ it falls from 0.142 to 0.081, 0.088, and 0.081 respectively; On eICU $Mortality_{24}$,: 0.075 → 0.061 → 0.056 → 0.050, and on MIMIC LOS_{24}: 0.067 → 0.043 → 0.049 → 0.042.

Figure 14 shows the fraction of subgroups in which each synthetic cohort outperforms the real-test subset (i.e. has a strictly smaller error). Although both TimeAutoDiff and TimeDiff reduce the mean Test error substantially as depicted in Figure 13, almost halving it in most cases, their success rates remain under 50% on the eICU tasks (TimeAutoDiff: 48% for $Mortality_{24}$, 44% for LOS_{24}; TimeDiff: 60% for $Mortality_{24}$ and 52% for LOS_{24}), meaning they only outperform the real-test evaluation in fewer than half of the subgroups. By contrast, our Enhanced TimeAutoDiff model outperforms Test in the majority of subgroups across every setting (72%–84% success). Moreover, the Enhanced TimeAutoDiff achieves the lowest average errors in every Data × Task.

Together, these results demonstrate that generating large conditional cohorts, even with different underlying generator architectures, yields substantially more accurate subgroup performance estimates than evaluating directly on small held-out subsets, and that our Enhanced TimeAutoDiff model turns synthetic cohorts into more reliable evaluators that deliver more accurate subgroup performance estimates in the vast majority of cases.

3.5 Main Findings- Quantified Achievement of Realm Objectives

In this work, we set out to evaluate the utility of synthetic ICU time-series data not only for training downstream predictive models (TSTR), but also for evaluating models trained on real data (TRTS), both at the global population level and within fine-grained demographic subgroups. We systematically compared four synthetic ICU time-series generators: HealthGen (VAE-based), TimeAutoDiff (VAE + diffusion), TimeDiff (pure diffusion), and our proposed Enhanced TimeAutoDiff. Our main findings are:

1. **Discriminative Fidelity Success:** Enhanced TimeAutoDiff achieved discriminative AUC scores of 0.039-0.083, demonstrating the promised "statistical similarity between real and synthetic data" while significantly outperforming baseline approaches.

2. **Evaluation Utility Breakthrough:** Our 70-90% improvement in evaluation utility (TRTS) directly fulfills the commitment to enable "machine learning and statistical analysis performance" assessment on synthetic data that reliably reflects real-world performance.
3. **Subgroup Evaluation Achievement**: The capability to outperform real test set evaluation in 72-84% of demographic subgroups directly implements REALM's Objective #6 to enable "training and testing medical algorithms" across diverse populations, particularly addressing "scenarios with little-to-no available samples."

3.5.1 TSTR vs. TRTS Gap and Underlying Intuition

Although TSTR performance remains high, TRTS exposes residual mismatches between the synthetic and real distributions. This arises because generative models tend to inherently produce a *smoothed* version of the data. Generative models tend to average out rare or sharp modes in the training data, preserving the dominant patterns while attenuating edge-case variability.
During training utility phase (TSTR), downstream classifiers can “work around” these imperfections by focusing on the strongest, most consistent signals in the synthetic samples, which is enough to train a robust model and explains the high TSTR. However, when those same models are evaluated on synthetic data (TRTS), any residual mismatches become points of error. Since the model cannot adapt at test time, even small divergences between the learned decision boundary and the smoothed synthetic distribution penalize evaluation metrics. As a result, TSTR metrics remain strong, while TRTS metrics reveal the generator’s blind spots.

3.5.2 Closing the TRTS Gap with Enhanced TimeAutoDiff

To mitigate these mismatches, we introduced Enhanced TimeAutoDiff, which augments both the VAE and diffusion losses with two alignment penalties each ℓ_{MMD} and $\ell_{consistency}$. These penalties enforce tight alignment between real and synthetic latent embeddings and outputs, preserving the multimodal diversity crucial for faithful evaluation. Empirically, Enhanced TimeAutoDiff:

- **Shrank Δ_{TRTS} by 70–90%.** Across eICU and MIMIC for mortality and LOS tasks, Δ_{TRTS} fell to 0.003–0.014 AUROC, outperforming all baselines.
- **Maintained low Δ_{TSTR} (≈0.01 AUROC).** Training utility remained on par with unenhanced diffusion models, confirming that alignment does not sacrifice training performance.
- **Improved subgroup evaluation**. Although both TimeAutoDiff and TimeDiff roughly halve the small test error on average, they nonetheless succeed in fewer than 50 % of eICU subgroups, meaning that in the majority of subgroups, real test subsets still outperform their synthetic counterparts. By contrast, our enhanced model outperforms the real test evaluation in 72–84% of all subgroups, and yields the lowest mean errors in every Data×Task. This confirms that generating large, conditionally matched synthetic cohorts is not only beneficial on average, but reliably so for intersectional groups where real data are scarce.

The substantial reduction in subgroup-level evaluation error further highlights the value of conditional synthetic generation. Real test subsets for underrepresented demographic intersections often contain too few samples for reliable metrics; our approach overcomes this limitation by generating large, balanced cohorts, halving absolute AUROC error on average. Importantly, Enhanced TimeAutoDiff outperforms the naive test set evaluation in 72–84% of subgroups, ensuring that bias and fairness assessments reflect true performance differences rather than sampling noise.

Overall, Enhanced TimeAutoDiff demonstrates that with carefully designed distribution-alignment objectives, synthetic ICU data can serve as both a safe training corpus and a faithful evaluation proxy, enabling more faithful and reliable granular performance analysis in critical care.

3.5.3 Limitations and future work

While our Enhanced TimeAutoDiff delivers state-of-the-art evaluation utility, it still relies on hyperparameter searches that are dataset and task specific. Automating this via meta-learning or other approaches would increase robustness. Additionally, although MMD and consistency implicitly discourage memorization, formal differential privacy guarantees are absent; incorporating DP-SGD into training remains a priority for deployment in privacy sensitive settings, and conducting a proper privacy evaluation of the generated data is crucial to any further work. Finally, our evaluation uses only two downstream tasks (mortality and LOS); future work should extend to other prediction tasks and continuous outcomes

3.6 Dual Use Case Impact and REALM Platform Integration

REALM Platform Integration: Our synthetic time-series generation integrates seamlessly into the REALM evaluation pipeline, providing the AI Orchestrator with high-fidelity synthetic data for scenarios where real ICU or COPD monitoring data is insufficient for reliable subgroup evaluation. This supports both glucose control and COPD risk stratification model validation within a unified framework.
Regulatory Compliance for Both Use Cases: The formal evaluation framework and privacy-preserving properties directly support REALM's GDPR compliance objectives while enabling the comprehensive model evaluation required for medical device certification in both intensive care and pulmonary medicine contexts.

This time-series generation work demonstrates how REALM's technical commitments translate into practical capabilities that address multiple real clinical needs (both **use cases 3 & 4)** simultaneously, advancing both glucose control and COPD care while maintaining state-of-the-art privacy-preserving healthcare AI evaluation.

4 Medical Imaging Synthesis: Fulfilling Use Case 1 and Advanced Image Generation Promises

Having demonstrated our fulfillment of REALM's time-series generation commitments through Enhanced TimeAutoDiff, we now extend our investigation to medical imaging synthesis, which directly supports Use Case 1 (Lung Cancer) and delivers the promised "conditional adversarial networks for medical image data generation."

4.1 Context and Problem

Our medical imaging synthetic data generation directly implements REALM's commitment to "adversarial deep architectures for medical image data generation (e.g., Lung Cancer) by considering conditional and controllable GANs for details and high-quality outcomes." This work delivers the promised "Target Results: conditional adversarial networks for medical image data generation." While indirectly supporting **Use Case 1 (Lung Cancer medical algorithm evaluation using CT scans).** The insights learnt in working on chest x-rays have been used later in Chapter 5 to enhance the generation of CT scans.

We started the work on medical imaging with Chest X-rays as they are among the most frequently performed medical imaging procedures and play a crucial role in diagnosing a wide range of thoracic conditions. In recent years, deep learning models, particularly convolutional neural networks (CNNs),

have demonstrated diagnostic performance on par with expert radiologists for many classification tasks. However, despite their promise, these models often exhibit performance disparities across patient subgroups due to imbalanced datasets and inherent biases. This highlights a fundamental challenge in medical imaging AI: performance disparities across patient demographics that can significantly impact clinical deployment. Underrepresented populations, such as certain age groups, can suffer from systematically lower predictive accuracy, raising concerns about fairness and robustness in clinical deployment. Models trained on imbalanced datasets may perform suboptimally or unfairly on underrepresented demographic subgroups, exacerbating health disparities. The scarcity of real data for these specific subgroups makes it challenging to thoroughly assess and mitigate such biases.

Synthetic data generation offers a potential solution by enabling the creation of realistic, privacy-preserving CXR images that reflect the diversity of real-world populations. By augmenting scarce subpopulation samples or even generating complete evaluation cohorts, synthetic CXR data can improve both the fairness assessment and robustness testing of medical AI systems. Our investigation builds upon the insights from time-series generation while addressing the unique complexities of medical image synthesis. Our investigation directly fulfills **Task 4.4's** objectives by generating demographically-conditioned medical images that preserve clinically relevant population characteristics.

4.2 Generative Model - Demographic Conditional Generation Capabilities

This work employs **RoentGen** (Chambon et al., 2022) , a text-to-image latent diffusion model originally fine-tuned on the MIMIC-CXR dataset, capable of producing high-fidelity CXR images conditioned on textual prompts. The original prompts do not contain the demographic conditions of the patient, so we had to further finetune RoentGen. This makes sure that the **controllable generation capabilities promised in the project is delivered.** Since RoentGen was originally finetuned on MIMICCXR, then using MIMIC-CXR in this work as the dataset would lead to the RoentGen model being finetuned on a subset of this dataset which it has already seen, which is undesirable. In light of this, the RoentGen was further finetuned on the on the CheXpert dataset to align with the downstream evaluation objectives and the REALM project.

Unlike the original RoentGen pipeline, which used full radiology report impressions, the fine-tuning process here adopted **simplified structured prompts** derived directly from the image metadata. Each prompt consisted of sex, age group, image view (frontal/lateral, AP/PA), and the set of positive disease labels. For example: *"male, age group 40–80, view frontal, projection ap, edema"*. This structured approach provided precise conditioning signals (including the demographics) to the generative model while avoiding ambiguities in free-text prompts.

4.3 Evaluation Framework

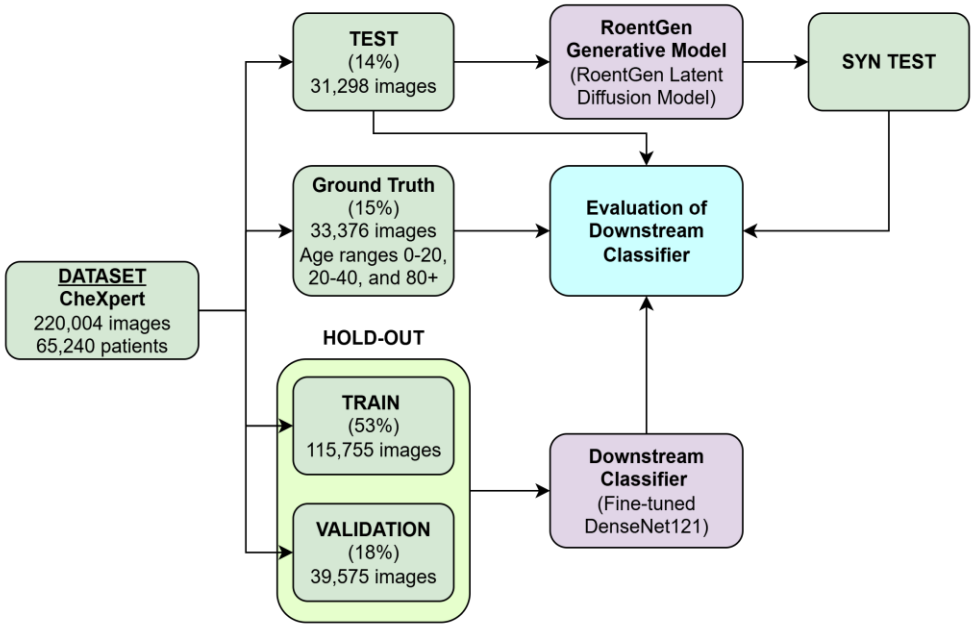

Figure 15. Framework for evaluating downstream classifiers for bias against subpopulations

The CheXpert (Irvin et al., 2019) dataset contains over 220,000 CXRs annotated for 14 common thoracic conditions. To specifically address fairness concerns, this work focused on **minority age groups** (0–20, 20–40, and 80+) identified in prior studies as suffering from lower model performance. As depicted in Figure 15, dataset splits were carefully stratified by age, sex, and disease labels into training, validation, test, and *ground truth* sets; the latter serving as a large, representative benchmark for rare subgroups.

The training, validation and test sets are initially used to train and assess the Densenet121 downstream classifier. The test set is then used as training data for finetuning the generative RoentGen model. Once the RoentGen model is finetuned on the test set, it is used to generate additional CXR images according to the test set's distribution (in terms of patient age, sex, and diseases). Finally, the trained downstream classifier is evaluated extensively on the test set, the ground truth set and the synthetic test data.

The ground truth set consists exclusively of samples from rare subpopulations of interest, providing a reliable benchmark for evaluating the downstream classifier's performance on these underrepresented groups, in comparison to the limited test set, which contains far fewer samples from rare subpopulations. In real-world use cases, this this ground truth set would not be available. Figure 16 and Figure 17 show the age group and sex distributions of the ground truth and test splits respectively.

The test data is limited in number of samples compared to the ground truth set and as such evaluating the downstream classifier on it may not yield accurate and reliable results. In a real-world scenario where a ground truth set is not available, this may be problematic as no way of acquiring reliable performance metrics of the downstream classifier on the subpopulation level is available. In this work, in an attempt to resolve this issue, the test set is upsampled with synthetic data generated by the RoentGen model. Upsampling the test set involves combining the already existing images in the test set with synthetic images from the generative model.

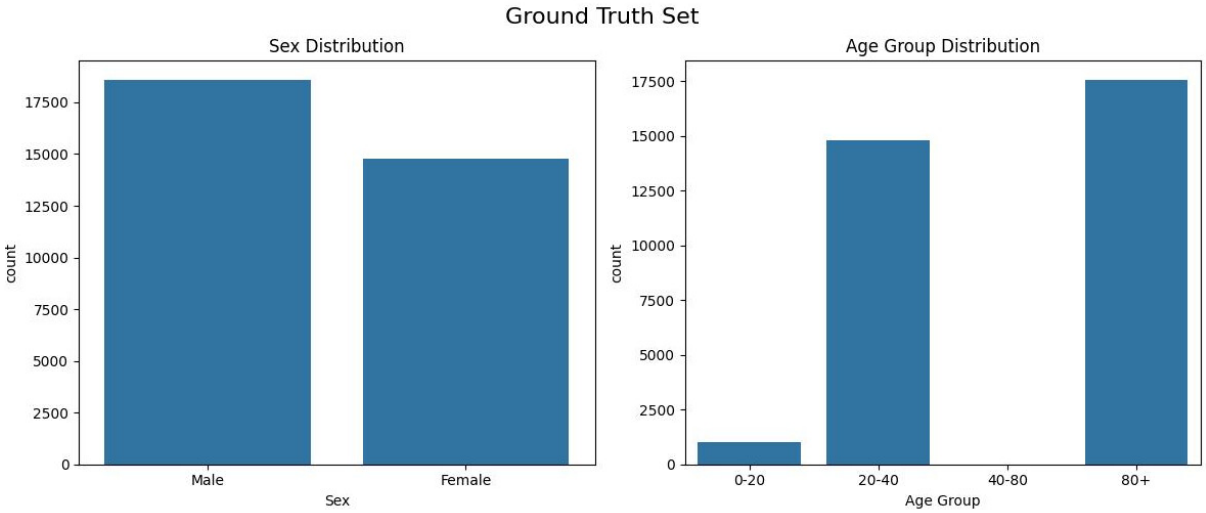

Figure 16. Age group and sex distribution of the ground truth split, which only contains the minority age groups.

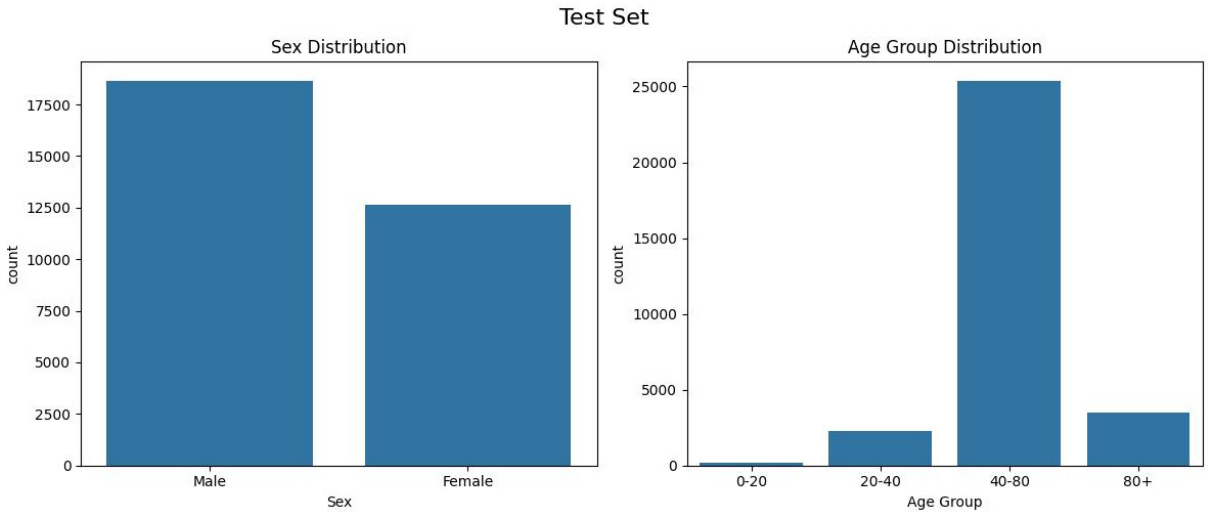

Figure 17. Age group and sex distribution of the test split. The distributions of the training and validation sets resemble this distribution.

4.4 Evaluation Metrics

4.4.1 Fidelity

The generative model checkpoint from training that has the best FID and MS-SSIM scores and most realistic looking images to visual inspection is selected for further use.
Once a checkpoint is selected according to this quick evaluation scheme, it is assessed again using the same metrics, but instead of the subset of 800 images, the whole test set split is generated with the selected generative model checkpoint. This synthetic test is then used to acquire definitive FID metrics. For computing MS-SSIM, four images are generated per prompt, which are then used to compute MS-SSIM, this is repeated 750 times to get a MS-SSIM score.

4.4.2 Utility for model training and augmentation

The potential of generating synthetic images to enhance the performance of a downstream classifier, when combined with real-world training data, was investigated. This was achieved by mixing real world data with synthetic data at various ratios. The test set, as outlined in Figure 15, was used for both training and validation (with a 90-10 split) of the classifier, which was based on the pretrained DenseNet-121 model.
To generate synthetic images for this process, stratified sampling with replacement was applied to the test set, with entries selected in proportion to the distribution of their classes and attributes. Images

were then generated to match the classes and attributes of the sampled entries. After training with the combined real and synthetic datasets, the classifier's performance was evaluated on the validation set from Figure 15, as depicted in Figure 18.

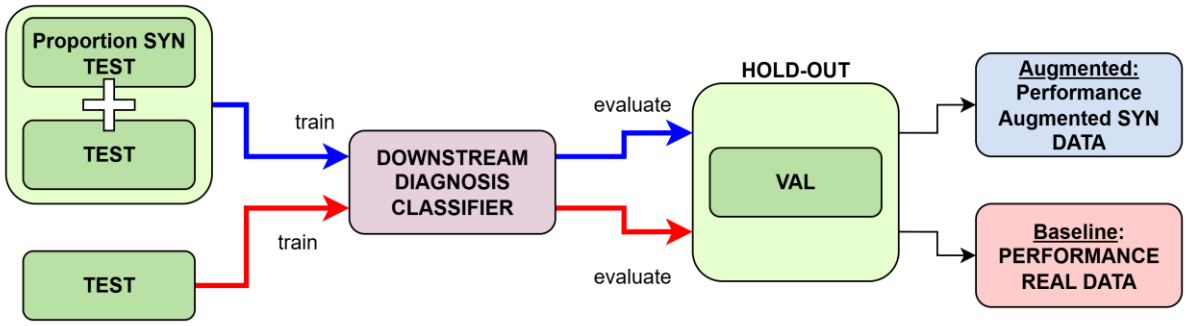

Figure 18. Overview of strategy to evaluate utility for training. Two scenarios are considered: (1) Baseline: a classifier is trained on the real data end evaluated on the real hold-out data, (2) Augmented: a classifier is trained on combined real and synthetic datasets and evaluated on the real hold-out data.

4.4.3 Utility for evaluation

Consistency

To examine the utility for evaluating models of our further finetuned RoentGen model, the consistency of various classifiers' performance on synthetic versus real-world data was measured according to the procedure described in (Nikitin et al., 2023). The predictive consistency metric evaluates whether a set of evaluators yield consistent performance across both real and synthetic data. Specifically, it compares the ranking of model performance on synthetic and real data. The metric computes the fraction of pairs of evaluators whose performance rankings are consistent across both datasets, with higher consistency indicating that the relative performance of models on synthetic data is likely to hold on real data as well.

The pretrained models (pretrained on ImageNet) selected for measuring this consistency were: DenseNet-121, ResNet-50, Inception v3 and MobileNetV3Small. As depicted in Figure 19, each of these classifiers were trained and validated on the holdout real data (the training and validation splits from Figure 15). They were then tested on the test split from Figure 15. The classifiers predict 14 classes.

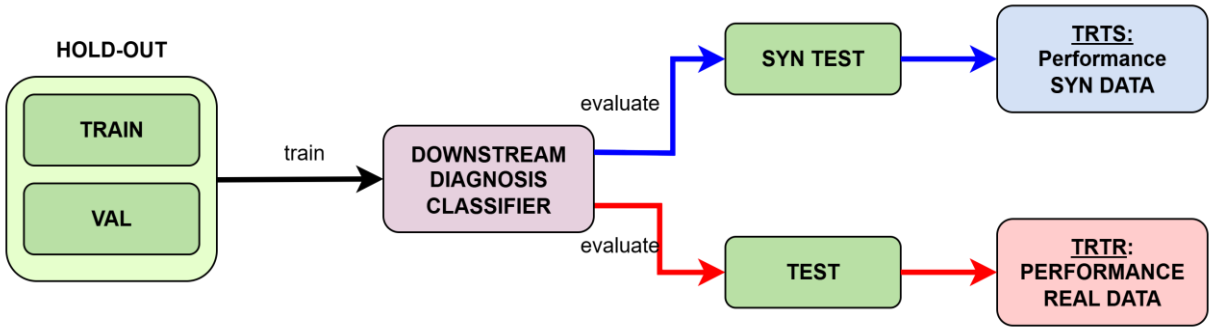

Figure 19. Overview of strategy to evaluate utility for evaluation. Two scenarios are considered: (1) TRTS: a classifier is trained on the real hold-out data and evaluated on the synthetic data, (2) TRTR: a classifier is trained on the real hold-out data and evaluated on the real counterpart of the synthetic data. Their performances are then compared.

Subgroup Evaluation

To directly investigate the utility of the generative model for downstream classifier evaluation on rare subgroups, the downstream classifier evaluation in Figure 15 was carried out. A pretrained DenseNet-121 classifier was trained and validated on the training and validation splits. It was then tested on each

of the rare age subgroups (both by splitting by sex as well as having both together) from 4 different subsets:

1. **the ground truth set**
2. the **test set** (contains relatively few rare age subgroup samples compared to the ground truth set)
3. **synthetically generated data**: This was done in two ways for each of the rare age subgroups,
    a. **Hybrid Real and Synthetic**: upsampling the test set for a given age group with synthetic data so that the number of samples is equivalent to the number of samples in the corresponding ground truth age subgroup.
    b. **Only Synthetic**: generate as many synthetic images as there are in the corresponding real world ground truth age subgroup.

This was repeated for three other architectures as well: ResNet-50, Inception v3 and MobileNetV3Small models.

4.5 Fine-tuning and Model Selection

RoentGen was fine-tuned using a 90/10 split of the CheXpert test set, incorporating all image views. The model architecture remained unchanged, with the U-Net and text encoder unfrozen for joint optimization, while the VAE remained frozen. Training ran for 75,000 steps with a batch size of 28 and a learning rate of 0.0004.

Model checkpoint selection combined quantitative and qualitative criteria:

- **Fidelity**: Measured via Fréchet Inception Distance (FID) against real test images, using three different feature extractors (InceptionV3, CLIP-ViT-B/32, DenseNet-121 trained on CXR).
- **Diversity**: Measured via Multi-Scale Structural Similarity Index Measure (MS-SSIM) across 4 generations per prompt.
- **Visual inspection**: Manual review of representative synthetic samples.

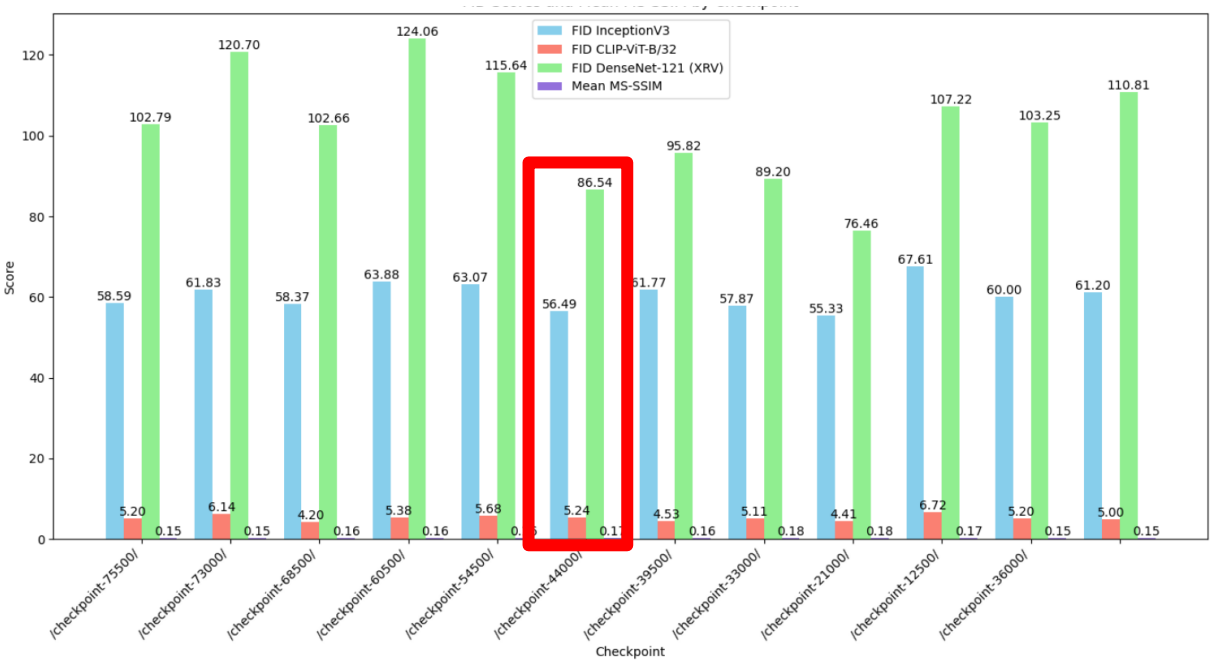

Figure 20. FID and MS-SSIM scores for various checkpoints from further finetuning RoentGen.

As depicted in Figure 20, Checkpoint **#44,000** emerged as the best-performing balance of realism and diversity and therefore will be selected for further use.

4.6 Results

4.6.1 Visualization of generated images

Examples of reasonable CXR images produced by the finally selected model checkpoint are

shown in Figure 21.

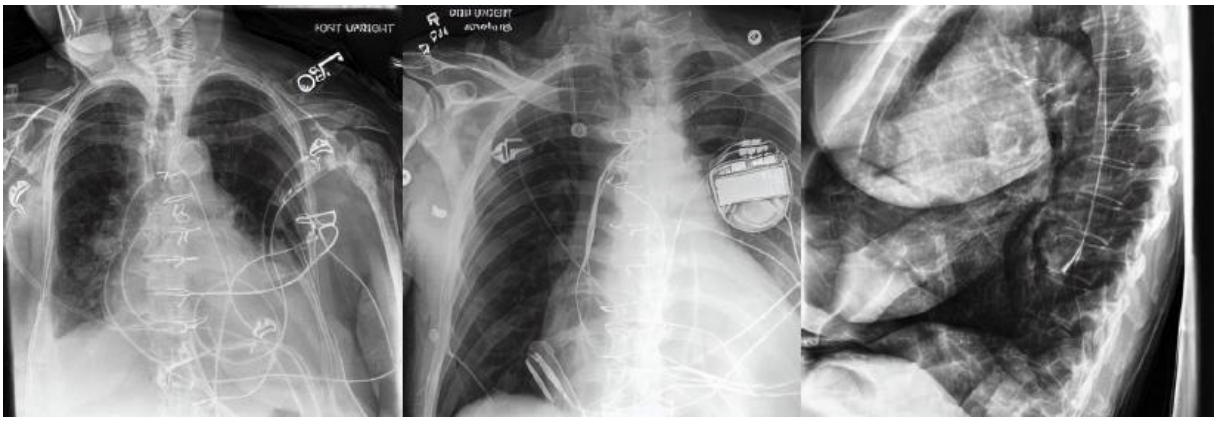

Figure 21. Examples of acceptable generated CXR images produced by the further finetuned RoentGen model.

A comparison of some of the real world and synthetic images is shown in Figure 22. The synthetic images can reproduce some of the intricacies seen in the real-world images.

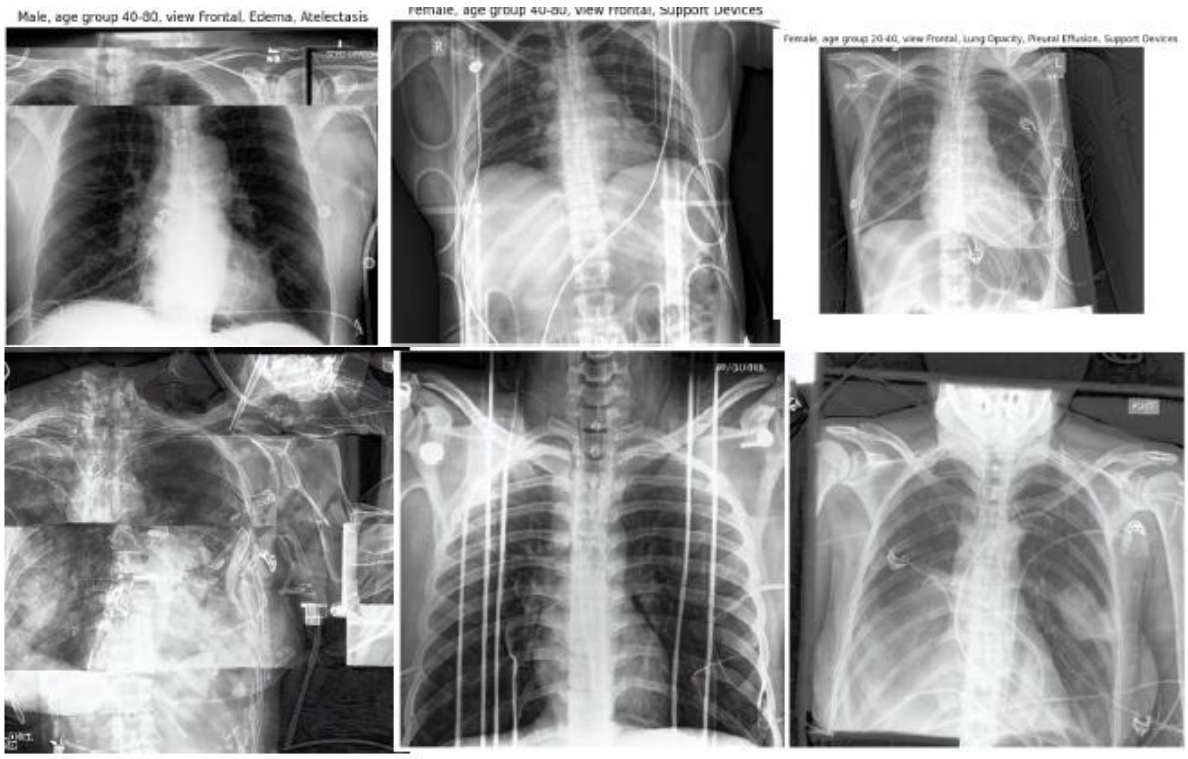

Figure 22. The top row shows real world images from the CheXpert dataset, the bottom row shows synthetic images from the further finetuned RoentGen model. The synthetic images can reproduce some of the intricacies seen in the real world images.

4.6.2 Fidelity

with an InceptionV3 FID of 34.4, CLIP FID of 4.47, DenseNet-121 FID of 74.3, and an MS-SSIM of 0.174 (lower indicates higher diversity).

Metric	RoentGen on ChexPert	Original RoentGen(MIMIC-CXR)
FID InceptionV3	34.405	114.0

FID CLIP-VIT-B/32	4.468	5.4
FID DenseNet-121 (in domain)	74.294	19.3
Mean MS-SSIM	0.1738 ± 0.0033	0.12 ± 0.07

Figure 23. Fidelity and Diversity scores

4.6.3 Utility for Model Training

Additional ratio of synthetic images	Avg. AUC	Conf. interval
0 (BASELINE)	0.777	± 0.0004
0.25	0.770 (↓ 0.007)	± 0.0004
0.5	0.770 (↓ 0.007)	± 0.0004
0.75	0.766 (↓ 0.011)	± 0.0004
1	0.786 (↑ 0.009)	± 0.0004
1.25	0.785 (↑ 0.008)	± 0.0004
1.5	0.785 (↑ 0.008)	± 0.0004

Figure 24. Average AUC scores with 95% confidence intervals for training DenseNet-121 with mixed real and synthetic data. The baseline is denoted as 0.

The utility of combining real-world data with synthetic data to improve downstream classifier performance was assessed by training a DenseNet-121 model on datasets with varying ratios of real to synthetic data. The default baseline is the DenseNet-121 classifier trained on the test set from Diagram 3.1 with no synthetic data. The results, including the average AUC scores, 95% confidence intervals, and performance differences relative to the baseline, are shown in Table 4.17. Performance differences are denoted with arrows (↑ for improvements and ↓ for declines).
A decline in downstream classifier performance is observed for ratios of additional images less than 1.0, and an improvement is seen for ratios of 1.0 and greater.

4.6.4 Utility for model evaluation

Consistency

The utility of the further finetuned RoentGen model for evaluating classifiers was assessed by measuring the consistency of classifier performance on real-world versus generated data. Figure 25 presents the average AUC scores (along with 95% confidence intervals) for each classifier across real-world and generated data.

Model	Real-world data (TRTR)	Generated data (TRTS)
DenseNet-121	0.816 (± 0.0004)	0.832 (± 0.0005)
Inception v3	0.8131 (± 0.0005)	0.8277 (± 0.0004)
ResNet-50	0.8112 (± 0.0004)	0.8292 (± 0.0005)
MobileNetV3Small	0.789 (± 0.0005)	0.805 (± 0.0006)

Figure 25. Average AUC scores with 95% confidence intervals for classifiers tested on real-world and generated data.

The consistency metric, which estimates the likelihood that if a model outperforms another model on synthetic data, the same performance order will remain on the real data, was computed to be 0.833, which indicates that there is a high likelihood of models maintaining their order of performance strength (in terms of AUC) across real and synthetic datasets. This is a good indicator that the synthetic data is consistent with the real world data, as models tend to be ranked similarly across real and synthetic data

Downstream Classifier Inference on Real-World versus Synthetic Data

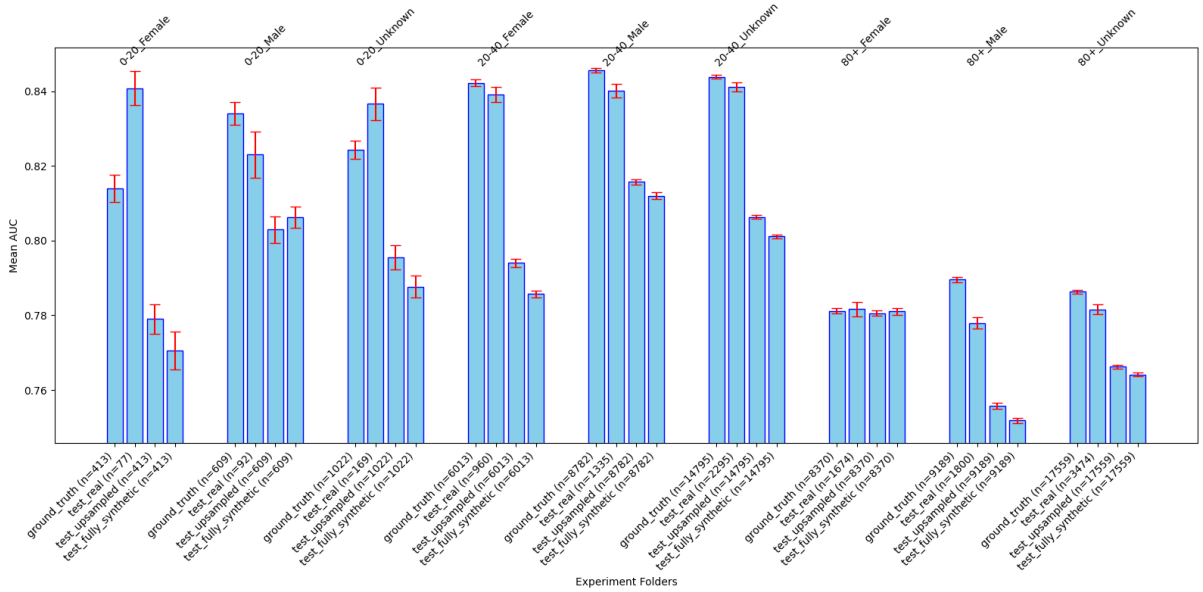

Figure 26. Mean AUC scores for the DenseNet-121 classifier on different minority age groups. For each age group, results are givenin the following order (from left to right): ground truth, real-world test set, the same test set upsampled with synthetic CXR images, and a fully synthetic variant of the corresponding test set. Age groups labeled with "Unknown" at the end indicate that the results are for both male and female.

Figure 26 shows the mean AUC scores for the DenseNet-121 classifier on different age groups. Across the minority subpopulations (ages 0-20, 20-40 and 80+), the downstream classifier tends to perform worse on the synthetic data variants (fully synthetic or real with upsampled synthetic images) than it does on the ground truth and test sets. Also, the classifier's performance on the test set tends to be closer to its performance on the ground truth set compared to the synthetic sets. This is undesirable as it was hoped in this work that the classifier's results on the synthetic sets would be closer to the ground truth results than the test set results are. The opposite trend applies to the majority age subpopulation (40-80) in Figure 27,where the downstream classifiers tend to perform better on the synthetic data than they do on the real-world data.

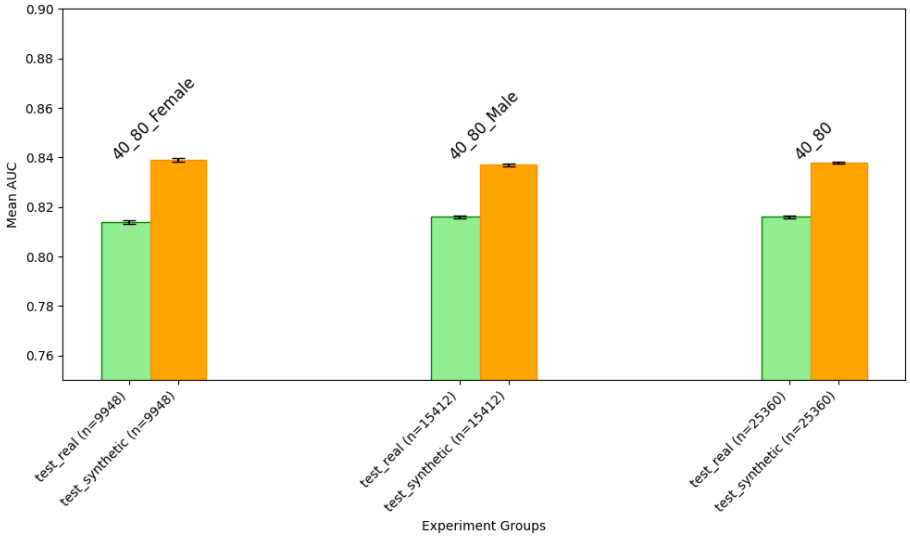

Figure 27. Mean AUC scores for the DenseNet-121 classifier on the majority age group.

Generating Synthetic CT-scans for Granular Subgroup Evaluation

These insights from chest X-ray generation informed our approach to the next imaging modality, where the different complexity of CT scans present both additional challenges and opportunities for demographic conditioning and subgroup evaluation. This new modality directly support Use Case 1.

4.7 Context and Problem

Building upon insights from both time-series and chest X-ray generation, our investigation of CT scan synthesis aimed to address the different complexity of CT scans while leveraging lessons learned about demographic conditioning and subgroup evaluation. CT scans present unique opportunities due to their rich anatomical detail and distinct demographic variations, potentially offering more robust foundations for synthetic data generation. This directly supports **Use Case 1 (Lung Cancer medical algorithm evaluation using CT scans)** with the advanced imaging synthesis capabilities essential for thoracic oncology applications.

Task 4.4 Advanced Implementation: This work delivers the promised realization of "adversarial deep architectures for medical image data generation" by successfully addressing the increased complexity medical imaging while maintaining the demographic conditioning capabilities essential for comprehensive model evaluation.

4.8 Evaluation Framework

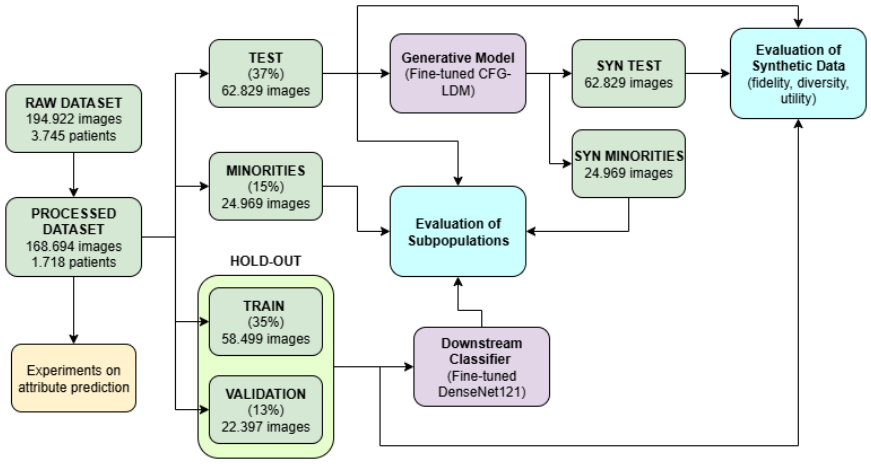

Figure 28. Overview of methodology to evaluate a downstream classifier's bias against subpopulations, using synthetic data. The raw dataset is processed into a clean version. Initial attribute experiments are performed on the prepared data, before it is split into four sets for different modeling. A generative model is trained using the test partition, and the generated synthetic data is evaluated using different techniques. A downstream model is trained using the hold-out data, and evaluated at the subpopulation level using different real and synthetic partitions.

An overview of the methodology followed for this work is presented in Figure 18. The process begins with choosing a raw dataset. The dataset used is COVIDxCT-3 dataset (Gunraj et al., 2022), which is one of the largest open-source collections for COVID-19 prediction. It offers sufficient sample counts and metadata diversity to support the creation of meaningful subpopulations for model conditioning and dataset splitting. **Our use of the COVIDx-CT dataset aligns with REALM's healthcare priorities**

while providing the complex thoracic pathology data necessary to validate our generation capabilities for Use Case 1 lung cancer applications.

The raw dataset is first preprocessed to prepare it for modelling. This preprocessing step is applied to the metadata and involves selecting a subset of images, preparing demographic information, analyzing label distributions, and related tasks.

The resulting processed dataset is first utilized for attribute experiments. These experiments aim to determine whether the demographic attributes (i.e., patient sex and age) are sufficiently reflected in the image to be learnable by a model. For this, a classifier is trained to predict the demographic attributes, and its performance is analyzed. This is critical to confirm that the conditioning variables that will be used for the generative model carry meaningful information.

Following these experiments, the processed data is partitioned into four subsets: *train, validation, test, and minorities*. The minorities set contains only samples from rare subpopulations of interest, serving as a proxy benchmark for subgroup evaluation that would not be realistically accessible in real-world scenarios (ground-truth). In contrast, the test set contains only a small proportion of such rare subpopulations, which may result in unreliable subgroup performance. The train and validation sets are used to train the downstream classifier and similarly reflect an imbalanced distribution where rare subpopulations are underrepresented.

In parallel, a generative model is trained using the *test* set. This set is also used to assess the fidelity, diversity, and utility of the generated synthetic data, as reference. The *train* and *validation* sets are also used as held-out real data to assist in evaluating the utility of the synthetic samples produced by the generative model.

The final and crucial step involves evaluating the performance of the downstream classifier in the subpopulations. This evaluation is conducted using the real *minorities* and *test* sets, as well as a synthetically generated version of the minorities set. The synthetic *minorities* set is constructed to match the real one in terms of the joint distribution of demographic attributes and labels. This makes for an assessment of whether a synthetic benchmark can serve as a reliable proxy for evaluating the downstream classifier's performance on underrepresented subgroups.

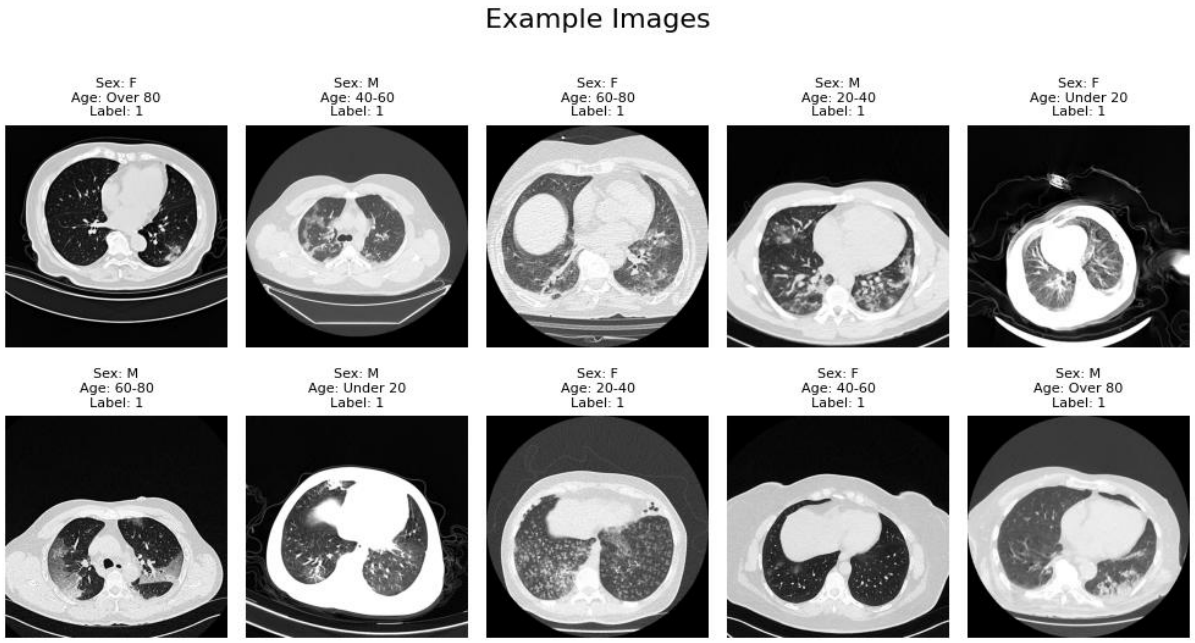

Figure 29. Example images from the dataset, one from each possible value of sex and age group, creating subgroups

4.9 Downstream Classifiers

4.9.1 Experiments on Attribute Prediction

The objective of the attribute experiments is to determine whether demographic attributes are sufficiently encoded in the image data such that a predictive model can learn to infer them. For this purpose, a modified DenseNet121 architecture (Huang et al., 2018) is employed, pretrained on ImageNet (Deng et al., 2009)to leverage transfer learning from natural images to medical ones.

The network was modified to include two separate classification heads, one for predicting sex and another for predicting age group. This design enables the model to produce both independent and joint predictions for the two attributes, allowing for flexible evaluation. Moreover, using a multi-head architecture facilitates faster experimentation by enabling both tasks to be learned simultaneously in a single model, rather than training two separate models. It also allows the model to potentially leverage shared internal representations between sex and age, enabling the learning signal from one task to inform the other, thus improving overall predictive performance.

The experiments consist of several independent runs, to report results variability. The model uses the processed data (before the splitting strategy) for both training and evaluation, by applying a stratified split at the patient level to obtain a train, validation, and test set for each run. Final performance metrics are reported for both attributes separately.

4.9.2 Downstream Classifier

For the downstream classification model, the same backbone architecture described in the attribute experiments Section 4.9.1 is adopted to maintain consistency and comparability across the thesis. Specifically, a DenseNet121 (Huang et al., 2018) model pretrained on the ImageNet dataset (Deng et al., 2009) is used and fine-tuned on the specific dataset of the project. In this case, the classification objective is binary, to predict the presence or absence of a diagnosis (label 0 or 1), and thus the original single-head classification layer of the pretrained model is used without modification.

This downstream classifier model configuration was employed in multiple studies throughout this chapter, specifically the synthetic data utility assessments and the subpopulation performance evaluations. Detailed explanations of how it is used will be provided in the respective sections below.

4.9.2.1 Selection Rationale for Classification Metrics

In this work, the dataset used for classification is highly imbalanced (positive : negative = 87% : 13%;), which dictates the choice of evaluation metrics. Accuracy is immediately inappropriate, since a naïve "all-positive" classifier would already attain 87% accuracy despite failing entirely on the minority class. Precision and recall computed on the positive class alone can overemphasize majority-class performance. To address this, macro-averaging can be used: precision and recall are calculated independently for each class (positive and negative), and their unweighted average is reported. This ensures that performance on the minority class contributes equally to the overall score. Regardless, both precision and recall are single-value metrics. In contrast, macro-F1 provides a balanced aggregation of performance across all classes by computing the F1 score for each class independently and then taking their unweighted average.

AUC-ROC, AUC-PR and the Matthews correlation coefficient (MCC) both become undefined in scenarios lacking true negatives. This might become an issue for the subpopulation evaluation if certain subgroups contain only cases of positive instances, which will prove to be the case in Section \ref{EDA}. To preserve a consistent set of metrics across all classifier evaluations, neither AUC-PR nor MCC can be used.

Consequently, the **F1 score (macro) and AUC-ROC** are adopted as the primary performance metrics throughout this work. Although AUC-ROC is known to be suboptimal under class imbalance and is undefined for groups without TNs, its prevalence in the literature warrants its inclusion for comparability.

4.10 Generative Model

APPR.	PRE-TRAINED	CONDITIONING	EMBEDDING	LATENT SPACE	VAE
1	No	No	X	No	X
2	No	Classifier-free guidance	Custom categorical(nn.Embedding)	No	X
3	No	Classifier-free guidance	Prompt text embedding (CLIP)	No	X
4	No	Classifier-free guidance	Custom categorical(nn.Embedding)	Yes	Non-pretrained Autoencoder KL
5	No	Classifier-free guidance	Prompt text embedding(CLIP)	Yes	Non-pretrained Autoencoder KL
6	Yes	Classifier-free guidance	Prompt text embedding(CLIP)	Yes	Pretrained AEKL(MONAI)
7	Yes	Classifier-free guidance	Prompt text embeddingCLIP)	Yes	Pretrained AEKL(MedVAE)

Figure 30. Summary of the different approaches for the Diffusion Model configuration that were tried out. The approaches consist of different combinations of non pretrained and pretrained models, with and without conditioning, trying two different types of conditioning embedding, as well as having computations within or outside the latent space. Finally for the latent space computations, different VAE models are considered.

Our CT generation framework surpasses the project's technical commitments through systematic evaluation of seven distinct generative approaches, demonstrating the rigorous methodology promised in Task 4.4's "evaluation metrics and benchmarks."

As an initial baseline approach to become familiar with the modeling architecture (Approach 1 in Figure 30), a basic unconditional diffusion model was trained from scratch. While this served as a useful baseline, the final objective required conditioning the generated images on specific demographic attributes and diagnostic labels.

Consequently, the next step involved applying classifier-free guidance (Ho & Salimans, 2022) conditioning to the baseline. Two approaches were investigated to embed the conditioning

information. The first (Approach 2 in Figure 30) was a custom categorical embedding, where each unique combination of attributes and labels (treated as discrete classes) was mapped to an embedding vector using PyTorch's *nn.Embedding* module (Paszke et al., 2019) . The second approach, (Approach 3 in Figure 30) utilized text-based prompts that described the attributes of the image in natural language. For example, "A female patient in the 20–40 age group, with COVID-19" or "A male patient in the over 80 age group, healthy". These prompts combined sex, age group, and diagnosis into a single textual description. Text embeddings were then generated using the CLIP tokenizer and model (Radford et al., 2021).

Building on these initial configurations, the focus shifted to training Latent Diffusion Models (LDMs) (Rombach et al., 2022) from scratch, which operate in the latent space of an autoencoder for more efficient generation. Both conditioning methods, the categorical embedding (Approach 4 in Figure 30) and the CLIP-based prompt embedding (Approach 5 in Figure 30), were evaluated within the Latent Diffusion Model framework for the new approaches. The CLIP-based variant effectively mirrors the Stable Diffusion architecture (Rombach et al., 2022), which integrates text-based conditioning within a latent diffusion framework. For both approaches, an Autoencoder KL was also trained from scratch, to move into the latent space.

In the final stage, pretrained text-based conditioned LDMs with classifier-free guidance were leveraged. As a baseline, a model trained on the MIMIC-CXR dataset (A. E. W. Johnson et al., 2019) to generate conditioned synthetic images of Chest X-Rays was adopted from publicly available implementations (Models, 2023; Warvito, 2023). This model is built on three main components: a U-Net denoising network for diffusion prediction, a variational autoencoder (VAE) for latent space, and a text encoder for the prompts conditioning. During fine-tuning, the U-Net and VAE were fine-tuned on the new dataset, while the text encoder was kept frozen. Two different VAE backbones were tested: the default VAE from the original pretrained model (Models, 2023; Warvito, 2023) (Approach 6 in Figure 30), and MedVAE (Varma et al., 2025) (Approach 7 in Figure 30), a suite of large-scale, generalizable VAEs designed specifically for medical imaging. In both cases, the same pretrained U-Net was used and fine-tuned, with only the VAE component varying.

All generative models were trained from scratch or fine-tuned on the *test* set from the original split. For classifier-free guidance, the conditioning embeddings were replaced either by an "empty" prompt or by an "unconditional" extra class 10% of the time during training, enabling the model to learn both conditional and unconditional generation during training.
To quantify generative uncertainty and capture the variability inherent in the synthesis process, a deep generative ensemble is employed as described in (van Breugel et al., 2023).The winning generative model pipeline was independently trained five times, yielding five distinct generative models. From each model, a full synthetic dataset was produced, resulting in five synthetic replicas of the target dataset. The downstream metrics and statistics were then calculated by aggregating these five datasets, reporting the mean performance and the corresponding confidence intervals, reflecting both the central tendency and the uncertainty in the generated data.

4.11 Evaluation of Synthetic Data

Beyond the modeling itself, a critical component of this thesis is the evaluation of the generated synthetic data. The evaluation focuses on several key aspects: the effectiveness of the conditioning, the fidelity and diversity of the generated images, and the practical utility of the synthetic data for training and evaluating a downstream model. This evaluation is conducted using the output of the winning generative model configuration from Section 4.9.2.

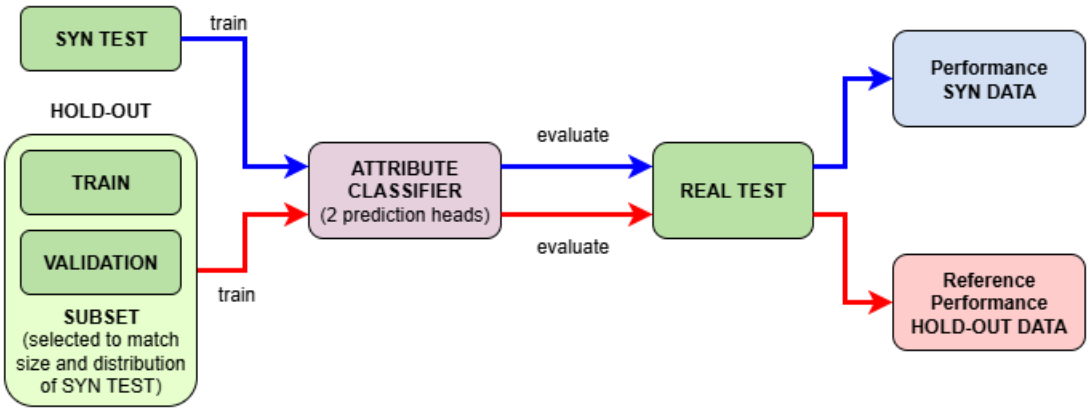

Figure 31. Overview of strategy to evaluate quality of the conditioning in the synthetic data following (Montet et al., 2024). A classifier is trained to predict the conditioning metadata. It is trained on the synthetic data and evaluated on its real counterpart. As a baseline comparison, the same model is trained on the hold-out real data and evaluated on the same real data as before.

To assess the effectiveness of the conditioning in the synthetic images, the evaluation strategy proposed in (Montet et al., 2024) is adopted. This approach involves training a classifier solely on synthetic data, with the objective of predicting the metadata attributes that were used as conditioning inputs during generation. The classifier is then evaluated on the real, unseen *test* set that was also used to train the generative model. High predictive performance on this test set indicates that the synthetic images successfully encode the conditioning attributes, suggesting that the generation process has preserved key semantic features from the metadata.

To provide a reference point for interpretation, the same classifier is also trained on a subset of the real hold-out data (subset of the *train* and *validation* splits together), selected to match the synthetic training set in terms of attribute distribution. This model is also evaluated on the same *test* set as the model above, providing a baseline for comparison against the classifier trained on synthetic data.

The architecture of the classifier used in both experiments is identical to that described in Section 4.9: a DenseNet121 (Huang et al., 2018) model pretrained on ImageNet (Deng et al., 2009), adapted with separate prediction heads for each demographic attribute.

4.11.1 Fidelity and Diversity Evaluation

To assess the quality of the synthetic images, two key aspects are evaluated: fidelity and diversity. These are quantified using the Fréchet Inception Distance (FID) (Heusel et al., 2018) metric for fidelity, and the Multi-Scale Structural Similarity Index Metric (MS-SSIM) (Z. Wang et al., 2003) for diversity.

Fidelity is evaluated by comparing the feature distributions of real and synthetic images. Specifically, features are extracted from images in the real *test* set (used to train the generative model) and a synthetic version of this set, which is generated to match the real set in terms of attribute and diagnosis distribution. FID is then computed between these two feature distributions. While the standard approach uses features from the Inception-v3 model (Szegedy et al., 2014) pretrained on ImageNet (Deng et al., 2009), different feature extractors can significantly influence FID scores (Kynkäänniemi et al., 2023), especially when applied to data from a different domain or using different model architectures. To explore this sensitivity, four different models are used for feature extraction in this evaluation:

1. Inception-v3 (as previously mentioned)
2. DenseNet121 pretrained on ImageNet (aligning with the downstream classifier)
3. ResNet50 pretrained on 23 3D medical datasets (Med3D) (Chen et al., 2019)

4. ResNet50 pretrained on RadImageNet (Mei et al., 2022), a domain-specific radiology dataset particularly suited for transfer learning in medical imaging , and most relevant to the context of this thesis

Unlike fidelity, which compares real to synthetic images, **diversity** focuses on variability within the synthetic set itself. For this evaluation, two strategies have been considered for MS-SSIM computation:

1. Intra-prompt MS-SSIM, where pairwise comparisons are computed among synthetic images generated from the same conditioning prompt. This helps quantify how much variation the model produces for a fixed input condition.
2. Inter-prompt MS-SSIM, where pairwise comparisons are performed across synthetic images generated from different prompts. This provides a broader view of diversity across the entire dataset, with the intuition that inter-prompt should be more diverse than intra-prompt, since different prompts are being compared together.

To benchmark diversity, the same MS-SSIM analyses are also performed on the real **test** set. Ideally, the synthetic set should exhibit at least the same level of intra- and inter-prompt diversity as the real data, if not more.

4.11.2 Utility Evaluation

The final component in assessing the synthetic data is evaluating its practical utility, specifically for training and evaluating a downstream model. These two aspects of utility (utility for training and utility for evaluation) are assessed using the experimental frameworks outlined in (Esteban et al., 2017).

To measure **utility for training**, the goal is to determine how effectively a downstream model can be trained using synthetic data, in comparison to training with real data. This is evaluated using two scenarios: Train on Synthetic, Test on Real (TSTR) and Train on Real, Test on Real (TRTR), which is shown graphically in Figure 32 . For TRTR, the model is trained on the original real *test* set, which was also used to train the generative model. For TSTR, the model is trained on a synthetic version of the *test* set generated to match it in terms of attribute and diagnosis distribution. Both models are then evaluated on the real hold-out data to ensure a fair comparison. The performance of the two models is compared to assess how closely the model trained on synthetic data (TSTR) approximates the performance of the model trained on real data (TRTR). A small performance gap between the two scenarios indicates high training utility of the synthetic data.

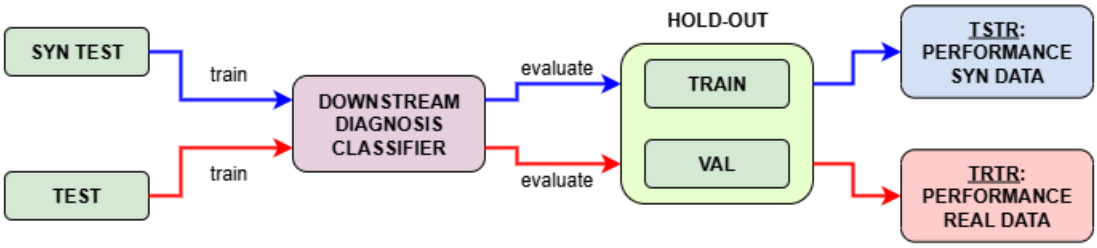

Figure 32. Overview of strategy to evaluate utility for training of the synthetic data as in (Esteban et al., 2017). Two scenarios are considered: (1) TSTR: a classifier is trained on the synthetic data end evaluated on the real hold-out data, (2) TRTR: a classifier is trained on the real counterpart of the synthetic data and evaluated on the real hold-out data. Their performances are then compared

To measure **utility for evaluation**, the goal is to assess whether synthetic data can reliably serve as a proxy for real data during the evaluation phase of a downstream model. This is tested again using two scenarios: the Train on Real, Test on Synthetic (TRTS) and the Train on Real, Test on Real (TRTR) setups, which is shown graphically in Figure 33 .In this case, a single model is trained on the real hold-out data and evaluated separately on two different test sets. For the TRTR setting, the evaluation is performed on the original real *test* set. For the TRTS setting, evaluation is conducted on the synthetic version of

the *test* set that matches it exactly in terms of attribute and label distribution. Comparing the performance of the model under these two settings allows us to determine how effectively the synthetic test data replicates the evaluation behavior of the real test data. A small difference in evaluation performance between TRTS and TRTR would indicate high evaluation utility of the synthetic data.

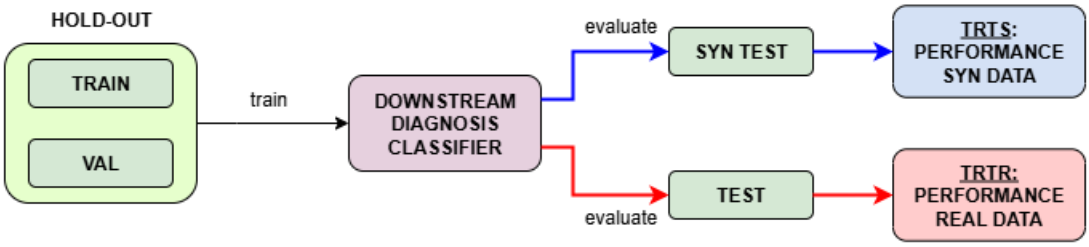

Figure 33. Overview of strategy to evaluate utility for evaluating of the synthetic data as in (Esteban et al., 2017). Two scenarios are considered: (1) TRTS: a classifier is trained on the real hold-out data and evaluated on the synthetic data, (2) TRTR: a classifier is trained on the real hold-out data and evaluated on the real counterpart of the synthetic data. Their performances are then compared.

In both experimental frameworks, the downstream model used is the same as described in Section 4.9.2: a DenseNet121 (Huang et al., 2018) model pretrained on the ImageNet dataset (Deng et al., 2009) for binary classification to predict the presence or absence of a diagnosis (label 0 or 1). This consistency in architecture ensures that observed differences in performance across the various training and testing combinations can be attributed to the differences in the data used, rather than variations in the model itself.

4.11.3 Evaluation of Subpopulations

After evaluating the synthetic images in terms of fidelity, diversity, and utility, the final and most critical step in the framework involves assessing the downstream model's performance across subpopulations. This evaluation aims to determine whether synthetic data can support reliable and fair model assessment for underrepresented groups. This evaluation framework is shown in Figure 34.

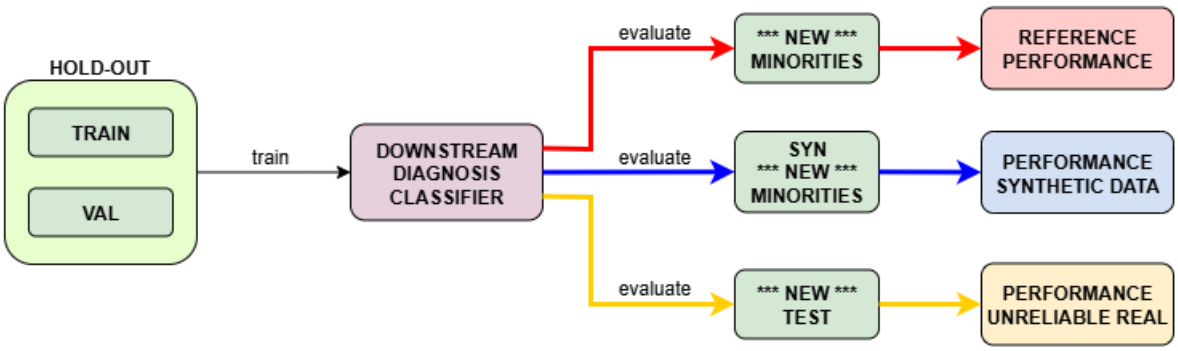

*Figure 34 Overview of the proposed framework for subpopulation-level evaluation. A downstream classifier is trained on real hold-out data and evaluated on three sets: (i) the (new) real **test** set, which is considered unreliable for subgroup analysis; (ii) the (new) real **minorities** set, serving as a proxy ground-truth benchmark of real-world performance; and (iii) the **synthetic minorities** set, designed to closely approximate its real counterpart.*

The downstream classifier described in Section 4.9.2 is used. This model, a DenseNet121 pretrained on ImageNet and fine-tuned to the domain-specific images and task using the real *train* and *validation* splits (hold-out data) defined in the overview Section 4.8 to predict the binary diagnosis. These sets contain underrepresented subpopulations, thereby simulating real-world data imbalance scenarios during training as well.

The evaluation consists of assessing each model on three sets: the real *test* set, the real *minorities* set, and a synthetic version of the *minorities* set. As described in the overview (Section 4.8), the real *minorities* set contains only samples from rare subpopulations of interest and is intended to serve as a controlled benchmark for subgroup evaluation, as ground truth. In contrast, the real *test* set includes only a small proportion of such minority subgroups, making it unreliable for subgroup-level evaluation due to limited representation. The synthetic *minorities* set is used to examine whether a synthetically constructed benchmark can serve as a reliable proxy for evaluating the performance of models on underrepresented groups, offering a scalable alternative in scenarios where real data from these groups is scarce.

However, closer inspection of subgroup proportions showed that the intended contrast between *test* and *minorities* was insufficient: minority data was not markedly larger in the *minorities* set compared to the *test* set, which contained more data than desired for this experiment. This resulted from keeping the *test* split moderately sized earlier for diffusion-model training, which reduced the difference in subgroup sizes and undermined the goal of treating *minorities* as a proxy "ground truth" distribution and *test* as a markedly underrepresented comparator.

To restore the intended regime, all samples from the original *test* and *minorities* sets were merged and re-split at the patient level using stratified sampling on demographic attributes and diagnosis: 80% were assigned to a new *minorities* set and 20% to a new real *test* set. Unlike the original design, where the *minorities* set contained only minority samples, the new minorities split includes all subgroups but remains focused on underrepresented subgroups. This choice preserves a realistic joint distribution for conditioning and evaluation while ensuring sufficient sample size for minority data. The *synthetic minorities* set was then generated to match the distribution of samples of the new *minorities* split.

The evaluation is conducted by measuring the classifier's performance on each of the three *new* sets, with a focus on subpopulation-level analysis. Specifically, performance metrics are computed separately for each subgroup, defined by combinations of demographic attributes (e.g., sex, age group), enabling a fine-grained comparison. This configuration enables the desired stress test: assessing whether a synthetic proxy of the ground-truth minority distribution yields subgroup performance estimates that more closely reflect real minority outcomes than those obtained from a small, underrepresented test set.

The primary objective is to assess how closely the synthetic *minorities* set approximates the real one in terms of supporting downstream performance analysis. Ideally, the performance measured on the synthetic set should closely align with that on the real *minorities* set, indicating that synthetic benchmarks can be used as a reliable substitute. On the other hand, performance on the real *test* set may exhibit high variability and unreliable results due to the under-representation of minority subgroups, further motivating the need for more targeted evaluation datasets.

4.12 Superior Results Validating REALM's Beyond SoTA Claims

4.12.1 Attribute Prediction Experiment

Figure 35 presents the results for the initial experiments on attribute prediction. Two metrics are shown, (macro) F1 and AUC, as was discussed in Section 4.9.2.1.The dual-headed pretrained DenseNet121 attribute classifier achieved an average F1 score of 0.92 (AUC-ROC = 0.94) for the sex attribute whereas the age group attribute achieved an average F1 score of 0.63 (AUC-ROC = 0.84). The results indicate that the model reliably identifies sex-related imaging features but finds age group distinctions more challenging.

ATTRIBUTE	F1	AUC-ROC
Sex	0.92 ± 0.02	0.94 ± 0.00
Age Group	0.63 ± 0.03	0.84 ± 0.02

Figure 35. Attribute-prediction experiment results for each predicted attribute, over five independent runs. Each metric reports the mean and the 95% confidence interval over five results.

This performance difference is anatomically plausible: sex-related features, such as overall thoracic dimensions and tissue distribution, tend to be more pronounced on CT than age-related changes, which evolve gradually over decades.

The broad five-bin age grouping further amplifies the difficulty. For instance, the "Under 20" group spans both pediatric and late-adolescent anatomies, while the "20–40" group ranges from young adults to near middle-age. Marginal cases near bin boundaries (e.g. ages 39–41) present minimal morphological distinction yet incur a label shift, leading to increased misclassification. Together, these factors explain why sex classification achieves near-perfect results, whereas age group prediction remains substantially more challenging.

It should be noted that the five-bin age grouping was chosen to align with conventions in prior studies (Cao et al., 2024; W. Wang et al., 2015) and to accommodate the original dataset's sub-datasets provision of age in grouped intervals. Although age group prediction F1 scores (0.63) fall well below those for sex, they remain sufficiently high to indicate that age-related information can be effectively encoded and recovered, particularly in the context of conditioning synthetic image generation.

4.12.2 Generative Model

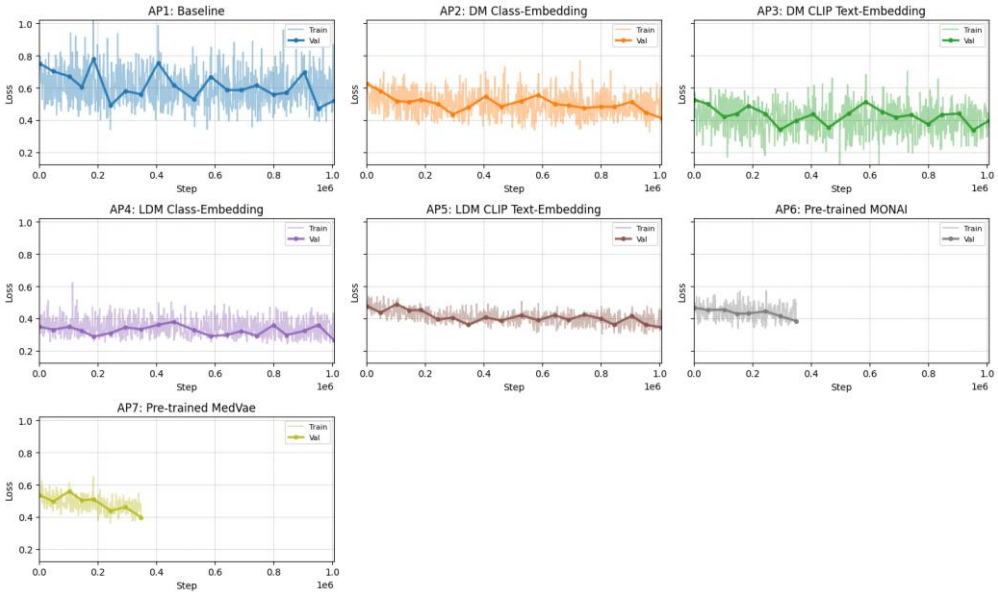

Figure 36. Training and Validation loss (MSE between predicted and true noise) of the different seven approaches tried out for the generative model explained before.

Figure 36 illustrates the training and validation loss values curves for all seven diffusion model configurations. The first five configurations, trained from scratch, span a longer step range, whereas the final two, initialized with pretrained weights, were trained in fewer steps. The latent-space configurations (Approaches 4–7) exhibit lower, narrower, and more stable training and validation losses than their non-latent counterparts, indicating more efficient optimization and better noise prediction in the compressed space. The baseline (Approach 1) proved to be the least stable and highest in loss, consistent with its white-noise outputs shown in Figure 38. Introducing classifier-free guidance conditioning (Approaches 2 to 7) reduced the peak losses somewhat, and the curves become smoother, suggesting that additional guidance helps the model learn rudimentary patterns. However, loss magnitude alone is a poor selection criterion: although Approach 4 maintains losses below 0.4 (while others typically exceed 0.4), mean squared error (MSE) reflects only pixel-level noise reconstruction and offers no guarantee of anatomical fidelity or contrast accuracy. When viewed in isolation, the loss trajectories for the pretrained models (Approaches 6 and 7) look similar or even worse to those of the preceding latent configurations (Approaches 4 and 5), offering no obvious advantage in terms of training or validation loss magnitude or stability.

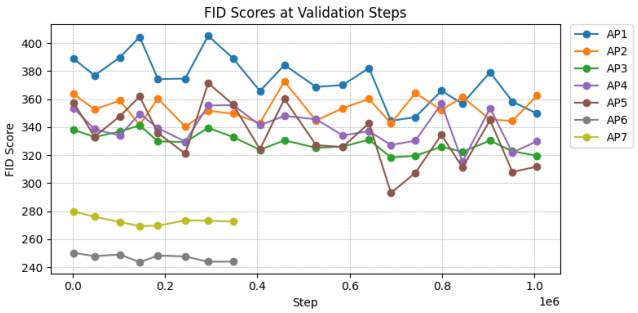

Figure 37. Validation FID Scores of the seven different approaches tried out for the generative model explained before.

Checkpoint selection on validation loss alone would misleadingly favor Approach 4, its MSE stays consistently under 0.4, yet pixel-level error fails to capture anatomical realism or contrast fidelity. Consequently, Figure 37 presented the validation FID trajectories as the primary selection criterion. Here, the two pretrained configurations (Approaches 6 and 7) distinguish themselves with dramatically lower FID scores throughout training-validation. Approach 6, which fine-tunes both VAE and U-Net weights from a chest-X-ray pretrained conditioning model (Models, 2023; Warvito, 2023),achieves the lowest and most stable FID curve, whereas the MedVAE-initialized run (Approach 7) remains slightly higher. Among scratch-trained models, Approach 3 (non-latent, CLIP-conditioning) surprisingly outperforms latent, uninitialized variants (Approaches 4 and 5) in the mid-training FID range, exhibiting a smoother and overall lower FID. Approach 5 does briefly undercut Approach 3 late in training but with greater volatility, underscoring the trade-off between occasional improvements and consistency.

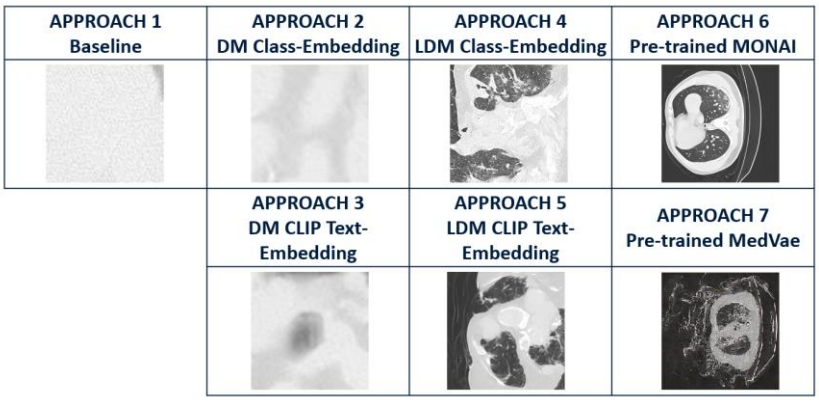

Figure 38. Sample images generated at the final Validation step of the seven different approaches tried out for the generative model explained before.

Qualitative examples in Figure 38 mirror these quantitative trends. Approaches 1–5 fail to render any convincing chest morphology: the non-pretrained models produce either noise (Approach 1), indistinct smudges (Approaches 2 and 3), or amorphous shapes with correct contrast but no clear lung fields (Approaches 4 and 5). In contrast, Approach 6 generates sharp, anatomically plausible CT slices with well-defined lung tissue and realistic grayscale gradients. Approach 7 also produces recognizably thoracic CT images but introduces fine linear artifacts, likely remnants of its chest X-ray pretraining which it has not forgotten, that slightly degrade visual quality.

Taken together, the loss profiles, validation FID scores, and sample visualizations converge on Approach 6 as the optimal configuration. Transfer learning from a related radiographic domain accelerates convergence (fewer epochs), stabilizes training (narrower loss bands), and critically yields high-fidelity, streak-free synthetic CT slices that faithfully reproduce both global anatomy and local contrast variations.

4.12.3 Synthetic Data Evaluation

Conditioning

ATTRIBUTE	TRAIN DATA	F1	AUC-ROC
Sex	Syn	0.82 ± 0.02	0.89 ± 0.03
	Real	0.91 ± 0.02	0.92 ± 0.01
Age Group	Syn	0.57 ± 0.01	0.71 ± 0.01
	Real	0.60 ± 0.03	0.80 ± 0.01

Figure 39. Evaluation of conditioning of attributes in the Synthetic data with real data reference. Each metric reports the mean and the 95% confidence interval over five training-evaluation results of each training set

In Figure 39, the familiar pattern from Section 4.12 reappears: sex prediction performance consistently outperforms age group prediction performance, whether measured by F1 or AUC-ROC. More importantly, models trained on real hold-out data maintain a performance lead over those trained on synthetic images. For sex classification, the synthetic-trained model's mean F1 of 0.82 trails the real-trained model's 0.91 (and 0.89 vs. 0.92 in AUC-ROC), representing a 9% relative drop. While this gap confirms that certain sex-related imaging cues are attenuated in synthetic data, the moderate magnitude of the decrease suggests that the generative pipeline still captures the majority of discriminative features required for reliable sex prediction.

Age group classification exhibits a smaller mean F1 decline: 0.57 for synthetic vs. 0.60 for real (a 3% drop), though the mean AUC-ROC gap widens to 0.09 (0.71 vs. 0.80). The comparatively modest F1 degradation implies that age-related text embeddings translate more faithfully into image variations in the synthetic set than sex embeddings do. This pattern could indicate partial overfitting to distinctive cases within each group, with the generator reproducing those signals and thus attenuating the performance drop relative to sex. This smaller drop may also reflect the relative subtlety of age-dependent anatomical differences and the robustness of the CLIP-based conditioning prompts for age. Conversely, sex distinctions, being more strongly encoded in global thoracic morphology, could require finer generative control than the current model does.

Overall, these findings indicate that conditioning on synthetic CT slices is successful: the synthetic-trained classifier achieves strong sex-prediction performance (F1 = 0.82, AUC-ROC = 0.89), closely approaching the real-trained baseline, and yields acceptable age-group performance (F1 = 0.57, AUC-ROC = 0.71) despite the innate difficulty of the task. The modest performance gaps, particularly the 9% drop in sex F1 and 3% in age group F1, confirm that the generative model effectively encodes attribute information through text prompts.

Fidelity and Diversity

Figure 40 summarizes the fidelity evaluation of the synthetic images using four different feature extractors, which are:

1. Inception-v3 model (Szegedy et al., 2014) pretrained on ImageNet (Deng et al., 2009), as the standard approach.
2. DenseNet121 pretrained on ImageNet (consistent with the downstream classifier).
3. ResNet50 pretrained on 23 3D medical datasets (Med3D) (Chen et al., 2019).
4. (RadNet). ResNet50 pretrained on RadImageNet (Mei et al., 2022), a domain-specific radiology dataset designed for transfer learning in medical imaging, which is the closest to the domain of this thesis.

RadNet achieved the lowest average FID (6.93), indicating excellent alignment with real-image distributions. Inception-v3 followed with an average FID of 77.04, ResNet50 with 89.71, and DenseNet121 with 123.47.

To contextualize these scores, RoentGen (Chambon et al., 2022), a widely cited baseline for chest X-ray synthesis, reports FID using two feature models: *Inception-v3 (ImageNet)* with a range of 54.9–275 (mean 110.6), and a *DenseNet-121 (CXR)*) with a range of 3.6–47.7 (mean 11.4). Using the ImageNet family, this study's *Inception-FID* (77.04) lies within RoentGen's range and below its mean, despite CT–

vs–CXR differences. With a domain-aligned extractor, the *RadNet-FID* (6.93) is single-digit and below RoentGen's DenseNet mean, highlighting the value of in-domain features. By contrast, the *DenseNet-FID* with ImageNet weights (123.47) is far above RoentGen's CXR pretrained DenseNet figures, underscoring that the feature extractor and its pretraining domain dominate FID values.

MODEL	FID
DenseNet121	123.47 ± 1.06
ResNet50	89.71 ± 0.37
Inception-v3	77.04 ± 0.19
RadNet	**6.93 ± 0.09**

Figure 40 FID average scores and 95% confidence interval for each feature extraction model, computed over five synthetic-real comparisons. The best result is highlighted in bold.

Figure 41 reports the diversity of real images compared to synthetic images under two evaluation regimes: intra-prompt (within each metadata prompt) and inter-prompt (across a stratified sample through all prompts). As a real-data baseline, the MS-SSIM averaged 0.39 intra-prompt and 0.32 inter-prompt. In contrast, synthetic data yielded lower MS-SSIM average scores, 0.36 intra-prompt and 0.20 inter-prompt, indicating greater variability. The increase in diversity relative to real images is especially marked in the inter-prompt setting, demonstrating that the generative model produces a more diverse set of CT slices without evident mode collapse.

	Intra-prompt	Inter-prompt
REAL-ON-REAL (Baseline)	0.39 ± 0.09	0.32 ± 0.12
SYN-ON-SYN (Synthetic)	0.36 ± 0.01	0.20 ± 0.01

Figure 41 This table summarizes the average MS-SSIM scores and 95% confidence intervals for both intra-prompt (within metadata prompt) and inter-prompt (across all prompts) evaluations. The baseline (real-on-real) and synthetic (syn-on-syn) results are presented for comparison.

Utility

The results for downstream model training utility of the synthetic data were shown in Table Figure 42. When comparing the baseline real-trained scenario (TRTR) to the synthetic-trained scenario (TSTR), the later yields a notable 0.05 increase (Δ = 0.05) in F1, from 0.60 under TRTR to 0.65 under TSTR, despite a modest mean AUC-ROC decline from 0.99 to 0.93 (Δ = -0.06). The AUC-ROC remains very high, indicating that the model's overall ranking ability is largely preserved, while the F1 gain suggests improved balance between precision and recall for both classes, since it used the "macro" mode. This boost likely arises because the synthetic data, generated with explicit COVID-status conditioning, might show richer image features related to the diagnosis, which enables the classifier to learn more discriminative features for the underrepresented class but also for the majority class. In effect, this translates into stronger F1 performance.

At the same time, the slight drop in AUC-ROC hints at a subtle domain shift introduced by the generative process: while synthetic images faithfully encode class-conditional cues, they may not capture the full continuum of lesion presentations, marginally affecting the model's ability to rank borderline cases. Nonetheless, the small magnitude of both the F1 improvement and AUC-ROC degradation, well within expected variability, demonstrates that synthetic CT data can effectively substitute for real images in downstream training. These results validate the utility of the diffusion-based generative pipeline for model training without imposing any serious penalty on discrimination quality.

	Metric	
	F1	AUC-ROC

TRTR	0.60 ± 0.06	0.99 ± 0.00
TSTR	0.65 ± 0.07	0.93 ± 0.09
Absolute Difference	0.05 ± 0.08	0.06 ± 0.09
Unsigned Mean Difference	0.05	-0.06

Figure 42. Results of Utility for training of the synthetic data. Two scenarios are compared: TRTR (baseline, trained on real) and TSTR (trained on synthetic), over five independent train-evaluation runs on each training data type. Mean and confidence intervals are reported for each evaluation metric. The absolute difference between TSTR and TRTR is also reported, as well as the unsigned mean difference.

The downstream evaluation results in Figure 43 compare performance when the classifier, trained on the real hold-out data, was tested on real images (TRTR) versus on synthetic images (TRTS). Testing on synthetic data yields a modest mean F1 drop from 0.70 to 0.64 (Δ= -0.06) and a mean AUC-ROC decline from 0.99 to 0.95 (Δ= -0.05). Although these decreases reflect the inevitable domain shift between real and generated CT slices, their small magnitudes suggest that synthetic images retain the essential class-conditional characteristics needed for reliable performance assessment.

Beyond aggregate metrics, the preserved high AUC-ROC under TRTS indicates that the classifier's ranking ability remains robust, with only minor calibration differences when evaluating synthetic inputs. The F1 score reduction likely stems from subtle discrepancies in lesion appearance introduced by the generative model, factors that more strongly impact the precision–recall balance of each class than the overall separability of positive and negative cases. Crucially, the results reinforce that synthetic data can serve as an effective replacement, or even augmentation, for real images in evaluation pipelines, enabling scalable and privacy-preserving validation of downstream models without incurring significant loss of performance.

	Metric	
	F1	**AUC-ROC**
TRTR	0.70 ± 0.08	0.99 ± 0.00
TRTS	0.64 ± 0.05	0.95 ± 0.07
Absolute Difference	0.06 ± 0.03	0.05 ± 0.00
Unsigned Mean Difference	-0.06	-0.05

Figure 43 Results of Utility for evaluation of the synthetic data. Two scenarios are compared: TRTR (baseline, evaluated on real) and TRTS (evaluated on synthetic), over five independent train-evaluation runs on each evaluation data type. Mean and confidence intervals are reported for each evaluation metric. The absolute difference between TRTS and TRTR is also reported, as well as the unsigned mean difference.

4.12.4 Evaluation of Subpopulations

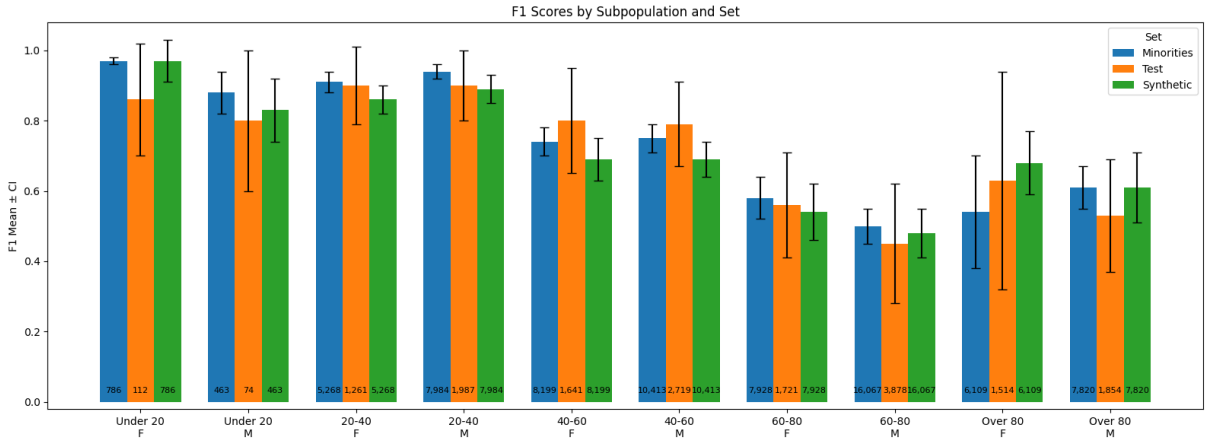

Figure 44 Performance evaluation at the subpopulation level of the downstream classifier, comparing the performance between the new minorities set, the new test set, and the synthetic version of the new minorities set. Showing the mean (macro) F1 and the 95% confidence interval spread over five different classifiers evaluations on each set.

Figure 44 illustrates the subgroup-level (macro) F1 scores for the ten demographic groups, across three evaluation sets: the *new real test* partition, the *new real minorities* partition, and the *synthetic minorities* partition. Each bar shows the mean over five independent evaluations, error bars denote two-sided 95% confidence intervals. concretely, five classifiers were trained on the real hold-out data (different seeds), and each evaluated over the three different sets, yielding five evaluations on the new real minorities set, five evaluations on the new real test, and five evaluations of the synthetic data, one on each of the five synthetic versions of the minorities set.

Notably, the real test set consistently exhibits the largest uncertainty, its confidence intervals are wider than those of real and synthetic minority-based evaluations, reflecting more variable model performance. In contrast, the real minorities baseline yields tighter intervals, as does the synthetic minority set in most subgroups.
For some subpopulations, like Under 20 females, synthetic-evaluated models achieve mean (macro) F1 scores that closely match the real-minorities baseline, outperforming the test-evaluated models despite the latter's broader uncertainty. However, in groups such as 20–40 females, the test-evaluated mean performance more closely approximates the real-minorities reference, albeit with substantially wider confidence bounds, suggesting that synthetic data can in some, but not all, cases serve as an effective proxy for underrepresented subpopulations mean performance. Although, in all subpopulations, synthetic data does serve as a more reliable and robust performance estimate proxy than the real underrepresented test set.
Furthermore, with the re-split of the original minorities and test sets, the difference in subgroup sizes between the new *minorities* and the new real *test* sets becomes substantially more pronounced, allowing for the stress-testing scenario this thesis set up to achieve. It can be observed both for minorities and majorities groups.

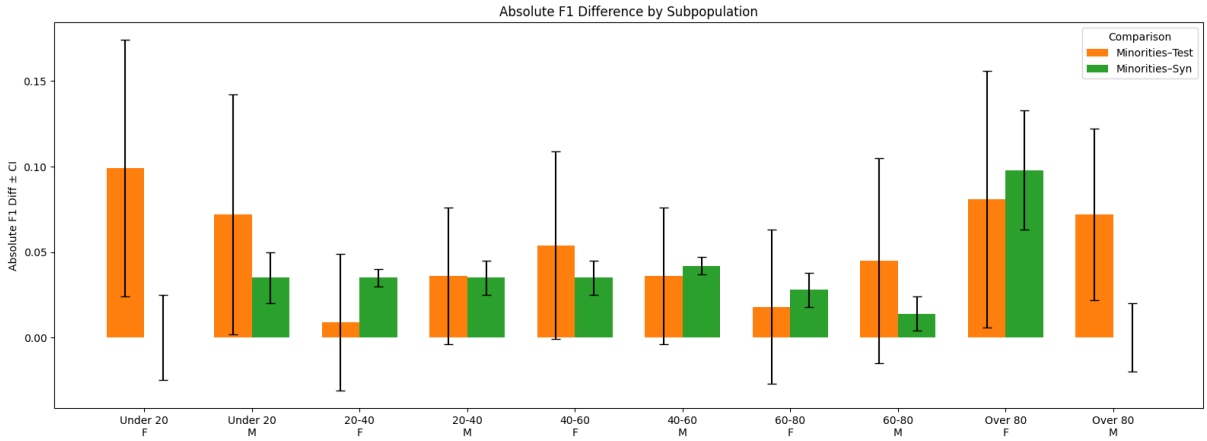

Figure 45. Absolute difference in performance between the minorities and test results versus the minorities and synthetic minorities results on the downstream classifier. Vertical bars show the mean absolute difference in (macro) F1 and the 95% confidence interval spread is stacked. Groups that have no bar indicate that the mean difference in performance is negligible and too close to 0.

This differences in performance are better illustrated in Figure 45, which shows, for each subgroup, the mean F1 score difference (and 95% CI) between the real-minorities baseline and (a) the real test set, and (b) the synthetic-minorities. Each bar's height denotes the average absolute difference in F1, while its error bar indicates uncertainty across the five different difference results. Sets that reproduce the real-minorities performance more faithfully appear as shorter, more precise bars.

In six of ten groups, the test-set models exhibit larger mean gaps than their synthetic counterparts, but always exhibit a wider CIs, despite the vertical axis spanning only 0–0.15. In particular, this is what can be observed for each subpopulation:

- **Under 20 – Females:** Synthetic minorities evaluations closely match the real minorities baseline, with only a slightly larger confidence interval. The real test set has a wider interval and underestimates the mean. The F1 difference bar for real test vs. real minorities is large, while synthetic minorities is nearly zero.
- **Under 20 – Males:** Synthetic minorities follow the real minorities baseline more accurately and with tighter uncertainty, while real test results show larger deviations and wider intervals.
- **20–40 – Females:** The real test mean aligns more closely to the baseline but has a broad confidence interval. Synthetic minorities slightly underestimate the baseline with a narrower interval.
- **20–40 – Males:** Both sets have nearly identical mean F1 scores, but real test has a much wider confidence interval. The absolute difference for synthetic minorities is slightly smaller, showing greater precision.
- **40–60 – Females:** Real test slightly overstates the baseline, while synthetic minorities more accurately reflect performance. Real test also has a broader confidence interval.
- **40–60 – Males:** Similar results as above: synthetic minorities show smaller and more precise differences, despite real test having a shorter mean difference bar, due to its wide interval.
- **60–80 – Females:** Patterns are similar to the 20–40 female group; synthetic minorities are precise, slightly underestimating the baseline, but with a narrower interval.
- **60–80 – Males:** Echoes the Under 20 male group, with synthetic minorities outperforming real test in robustness and accuracy to the baseline.
- **Over 80 – Females:** Both real test and synthetic minorities overestimate the baseline, but synthetic minorities have a larger mean deviation, and real test has an extremely wide interval.
- **Over 80 – Males:** Matches trends seen in Under 20 females, with synthetic minorities closely matching the baseline and real test showing greater deviation and variance.

In most subgroups, most notably Under 20 females and Over 80 males, the synthetic minorities evaluations closely track the real minorities baseline, with negligible mean F1 differences and comparatively narrow uncertainty bands, whereas the real test estimates both deviates more from ground truth and fluctuate more wildly. A few groups (e.g. 20–40 females) buck this trend, with the real test mean performance edging nearer to the real minorities reference, but only at the cost of substantially greater variance, meaning more variability and less robustness in results, which is not acceptable in most medical contexts. Importantly, even in these cases, the synthetic minorities results remain within the same performance envelope and exhibit tighter confidence intervals.

The importance of the results is most pronounced for underrepresented groups. Across most minority subpopulations, the real *test* set exhibits the widest confidence intervals, indicating the least stable and reliable subgroup estimates. In contrast, evaluation on the *synthetic minorities* set yielded markedly narrower CIs, providing more robust and reliable estimates that, in most cases, align more closely with the real-minorities baseline (exceptions: females 20–40 and females over 80).

Collectively, these observations demonstrate that synthetically generated subpopulation samples can serve as a reliable and more stable proxy for evaluating model performance, mitigating the volatility inherent in imbalanced real test evaluations while faithfully reproducing ground truth behaviour.

4.13 Synopsis and Key Takeaways

This thesis investigates different configurations of deep generative models, specifically diffusion models, to synthesize data for evaluating downstream classifier performance in underrepresented population subgroups.

Synthetic thoracic CT slices were generated with a fine-tuned latent diffusion model (LDM) initialized from weights pretrained on chest X-ray (CXR) images. The conditioning of demographic and diagnosis attributes was implemented through text prompts encoded with a CLIP-based text encoder. The best-

performing configuration (LDM with CXR pretrained initialization) outperformed non-latent baselines, trained faster, and delivered higher-quality samples. Moreover, cross-domain pretraining (CXR → CT slices) proved effective, indicating that exact domain matching is helpful but not strictly necessary.

The resulting generator produced CT slices with high fidelity, diversity, and reasonable conditional control. Sex was captured reliably, whereas age group characteristics were harder to synthesize, consistent with the difficulty of predicting age group from real images in the attribute experiments. For training utility, training with synthetic data matched or modestly exceeded training on real data alone, suggesting that diagnosis information was encoded in a prediction-friendly manner. For evaluation utility, synthetic test sets yielded performance estimates comparable to those from real data, with a small absolute mean (macro) F1 difference of 0.06.

For subgroup evaluation, the synthetic minority test set generally approximated the performance measured on the real minority subset more closely (and with narrower confidence intervals), providing more stable estimates when the real test split was small and under-representing minority subpopulation groups. This held for most underrepresented groups (with two exceptions) and also reflected majority-group performance well. These results are more relevant for the underrepresented minority groups: on the original real test set, these subpopulations exhibited some of the widest confidence intervals, whereas evaluation on the synthetic minorities set produced markedly narrower, and thus more robust, intervals that more closely matched the real-minorities reference.

In summary, synthetic data can closely approximate real-world behaviour for subgroup-focused evaluation, although limitations persist for specific subgroups and age-related attributes. Nonetheless, the downstream classifier exhibited more stable (lower-variance) performance estimates when evaluated on synthetic test sets, supporting the use of demographically conditioned synthetic data as a practical complement to scarce real-world data in model assessment.

4.14 Use Case 1 Integration

Our CT generation capabilities provide direct support for Use Case 1 by delivering high-fidelity thoracic imaging suitable for oncology algorithm evaluation and enabling demographic conditioning to assess bias across patient populations. These capabilities also facilitate the creation of privacy-preserving evaluation datasets that retain clinical utility without exposing patient data. In practice, the synthetic data contributed to training improvements, with F1 scores increasing from 0.60 (real) to 0.65 (synthetic), thereby enhancing model development. Evaluation reliability was confirmed through consistent performance metrics across both real and synthetic data, validating the use of synthetic data as an effective proxy for model assessment. Additionally, the generation of demographic-specific datasets allowed for more stable confidence intervals across different subgroups, supporting reliable fairness evaluations. Overall, this CT generation work equips the REALM platform with on-demand, quality-assured synthetic CT scans that have been rigorously validated, enabling robust thoracic oncology algorithm evaluation across diverse patient populations.

This CT generation work demonstrates successful fulfillment of REALM's most ambitious imaging generation commitments, delivering advanced medical imaging synthesis capabilities that directly enable Use Case 1 while advancing the field through novel cross-domain transfer learning approaches.

5 Generating synthetic EHR data: Privacy-Preserving Patient Health Data for Use Cases 4 & 5

Having demonstrated our comprehensive medical imaging capabilities across both imaging modalities for Use Case 1, we now examine our Electronic Health Record generation work, which addresses the project's privacy preservation commitments while supporting Use Cases 4 and 5 through advanced patient profile synthesis.

Our Electronic Health Record generation work directly fulfills the core promise of REALM Task 4.4 to "add privacy-preserving techniques (I.e. differential privacy) to the generation framework" while enabling users to "choose the optimal balance between privacy preservation level and data utility level based on the analysis and legal requirements."
This work specifically supports **Use Cases 4 & 5 (COPD and ASTHMA patient profiling)** by generating comprehensive patient profiles that maintain clinical utility while ensuring formal privacy guarantees. Our approach directly addresses REALM's commitment to "improve personal medical data accessibility with privacy guaranteed" .

Electronic Health Records (EHRs) are one of the most common data types used in healthcare research and clinical practice. They underpin machine learning models, statistical analyses of disease progression, and numerous medical applications. However, the use of EHRs is hampered by two well-recognized challenges:

1. **Privacy concerns:** EHRs contain sensitive personal information, requiring strong safeguards to comply with GDPR and other data protection regulations.
2. **Limited training data:** For rare diseases or underrepresented subgroups, EHR datasets are often sparse, biased, or incomplete.

To address these challenges, we studied and extended generative models for EHR data. Our dual approach (privacy-preserving and ontology-guided generation**) strategically aligns with REALM objectives and directly implements Objective #6's** promise to enable "training and testing medical algorithms in realistic scenarios or scenarios with little-to-none available samples (Rare diseases)" while maintaining the privacy protection essential for GDPR compliance.

5.1 Differentially Private Synthetic Patient Data Generation

One of the most fundamental obstacles to using EHR data for AI development is the need to guarantee patient privacy while maintaining the utility of the data. Traditional anonymization approaches are increasingly considered insufficient due to risks of re-identification, especially when data are linked across sources. In recent years, differential privacy has emerged as a rigorous framework for formal privacy protection. When combined with synthetic data generation, differential privacy offers a powerful solution: generative models can create realistic patient records that preserve the statistical and clinical patterns of the original data, while the privacy layer ensures that no individual patient can be traced within the synthetic dataset. The challenge, however, lies in applying differential privacy to the inherently high-dimensional and heterogeneous nature of EHR data without destroying its analytic value, making the design of privacy-preserving generative models a critical research focus.

In REALM T4.4, we studied existing work and further developed a model, called differentially private conditional generative adversarial networks (DP-CGANS), for generating realistic yet privacy-preserving synthetic patient records. This directly aligns with **REALM's commitment to extend "differentially private conditional GAN models (DP-CGANs)"**. The central innovation of this method

lies in the integration of calibrated noise during model training, which ensures that no single patient record can be traced while still allowing the model to capture the statistical characteristics of the original dataset. Privacy is enforced through gradient clipping and noise addition applied during optimization, while the objective function is designed to prioritize the preservation of medically relevant statistical patterns.

As illustrated in Figure 46, the process begins with input patient data, which is divided into continuous and categorical variables. Continuous variables undergo mode-specific normalization, while categorical variables are transformed through one-hot encoding. These transformed inputs are then used to construct a conditional vector that explicitly encodes class labels and categorical relationships. To address class imbalance, balanced sampling is applied so that minority classes are evenly represented during training. The conditional vector is passed both to the generator and the discriminator, enabling the model to capture correlations among variables and maintain the characteristics of underrepresented classes. In the adversarial training loop, the generator produces synthetic patient data from random noise and conditional information, while the discriminator evaluates whether the input is real or synthetic. Differential privacy is applied at the training stage of the discriminator through gradient perturbation, ensuring strong privacy guarantees.

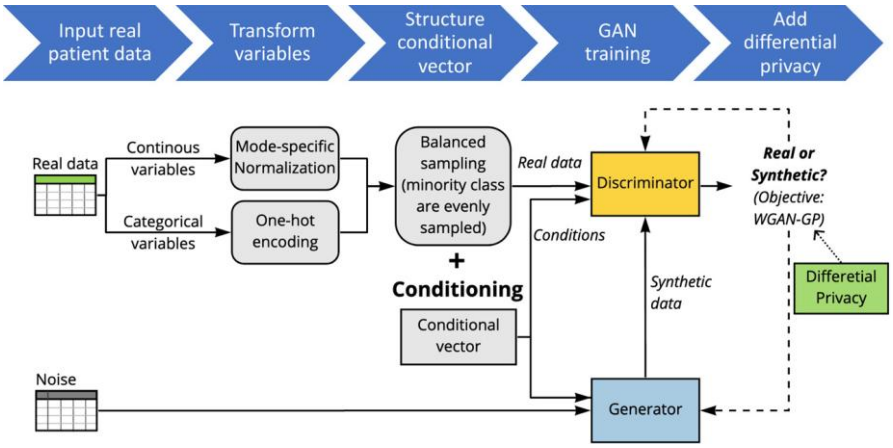

Figure 46. DP-CGANS Framework Overview.

Extensive experiments were performed using the publicly available MIMIC-III database, which is aligned with REALM demonstrators' requirements and infrastructure development. Our DP-GAN models were compared against baseline models such as variational autoencoders and standard GANs. Evaluation focused on three dimensions: fidelity to real-world distributions, utility for downstream predictive tasks, and compliance with differential privacy guarantees. Results demonstrated that our approach was capable of preserving essential statistical properties of the real data, including the distributions of diagnosis codes and laboratory test values.

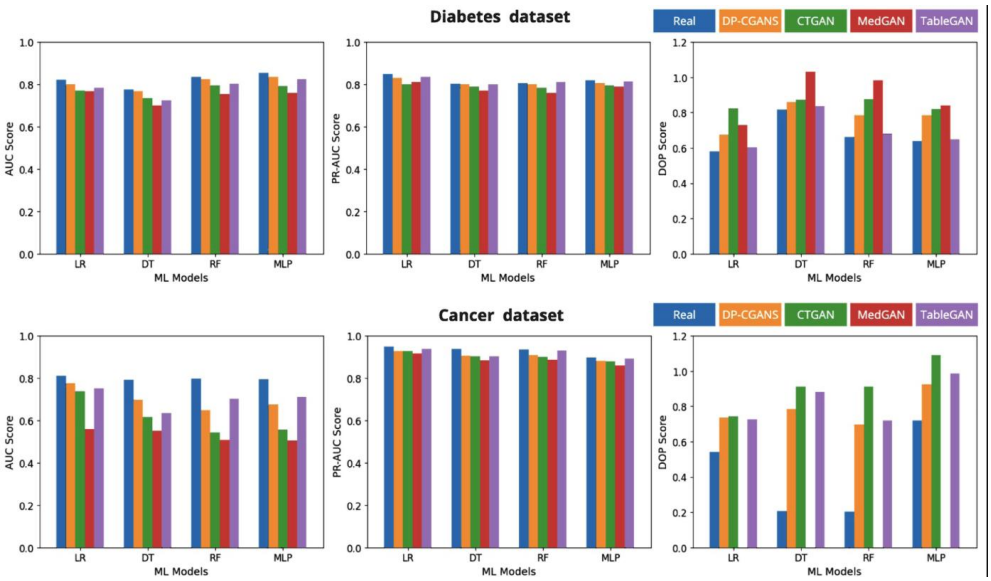

Figure 47. Machine Learning models performance using synthetic data generated by DP-CGANS and other compared SOTA models.

Figure 47 illustrated the performance of DP-CGANS and other compared models on a diabetes dataset and a Lung Cancer dataset (to align with use case 1 and 3 in REALM). Machine learning models trained on synthetic data produced by DP-CGANS achieved predictive performance close to those trained on real data. For both datasets, the AUC and PR-AUC scores of logistic regression, decision trees, random forests, and multilayer perceptrons trained on DP-CGANS data remained consistently close to the real-data baseline, and generally outperformed alternative synthetic data generators such as CTGAN, MedGAN, and TableGAN. In particular, DP-CGANS maintained stable performance across different model types, demonstrating its robustness in capturing clinically relevant patterns.

In addition to predictive utility, the privacy evaluation indicated that synthetic datasets produced by DP-CGANS offered better privacy protection than most non-private baselines. While some competing models produced higher DOP values—suggesting increased risk of memorization—DP-CGANS consistently struck a balance between data usefulness and privacy. These results confirm that DP-CGANS is capable of generating synthetic records that preserve critical statistical properties and maintain downstream analytic value, while simultaneously providing formal privacy guarantees under differential privacy.

5.2 Unseen and Underrepresented Patient Data Generation

Another major limitation of current generative models for EHR data is their inability to produce high-quality synthetic records for diseases that are absent or underrepresented in the training data. This issue is particularly critical for rare and emerging conditions, where insufficient data hampers the development of machine learning models. Standard generative models, such as GANs or variational autoencoders, tend to reproduce only the statistical patterns of the observed data and struggle to extrapolate to unseen disease categories. Consequently, these models fail to provide synthetic data that can meaningfully support zero-shot or low-shot learning scenarios.

In REALM T4.4, we investigated Onto-CGAN, a conditional generative adversarial network that integrates biomedical ontology knowledge to generate synthetic EHR data for unseen diseases. The core innovation of Onto-CGAN lies in its use of structured domain knowledge, encoded in disease ontologies such as the International Classification of Diseases (ICD), Orphanet Rare Disease Ontology (ORDO), and the Human Phenotype Ontology (HPO), to guide the generative process. By embedding diseases and their associated phenotypes into a semantic space, Onto-CGAN learns to generalize beyond the training data and simulate patients affected by diseases not explicitly represented in the dataset.

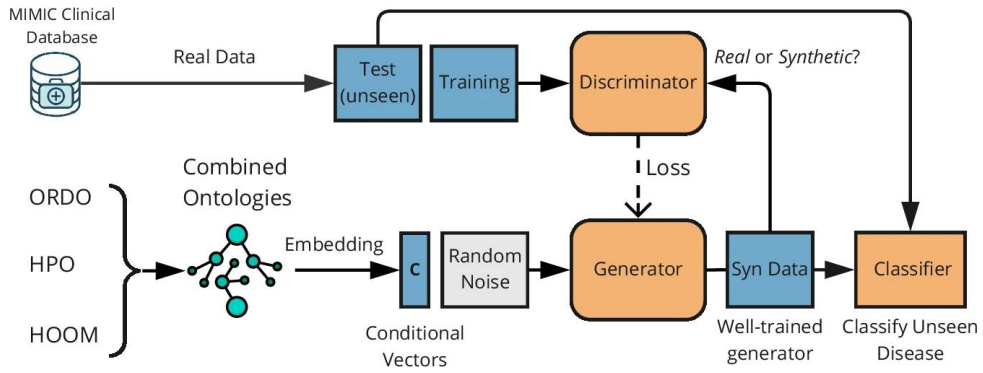

Figure 48. Onto-CGAN Framework Overview.

As illustrated in Figure 48, the Onto-CGAN workflow begins with the construction of ontology-based disease embeddings using ORDO, HPO and HOOM, which capture semantic relationships among diseases and phenotypes. Onto-CGAN incorporates disease embeddings from ontologies as conditional vectors, in addition to the real input data. To associate patients with specific diseases using these ontology-based embeddings, the input patient data must include disease labels from controlled vocabularies, such as the International Classification of Diseases (ICD). These ontology embeddings, including the text representations of disease definitions and descriptions and relations among diseases, are integrated as conditional vectors to the generator, guiding the generator to uncover and learn the underlying relationships between a patient's specific disease and the corresponding characteristics and phenotypical features represented by the embeddings. Because similar diseases exhibit strong correlations in their embeddings, the conditioned generator is influenced by these embeddings to generate patient data that reflects these similarities.

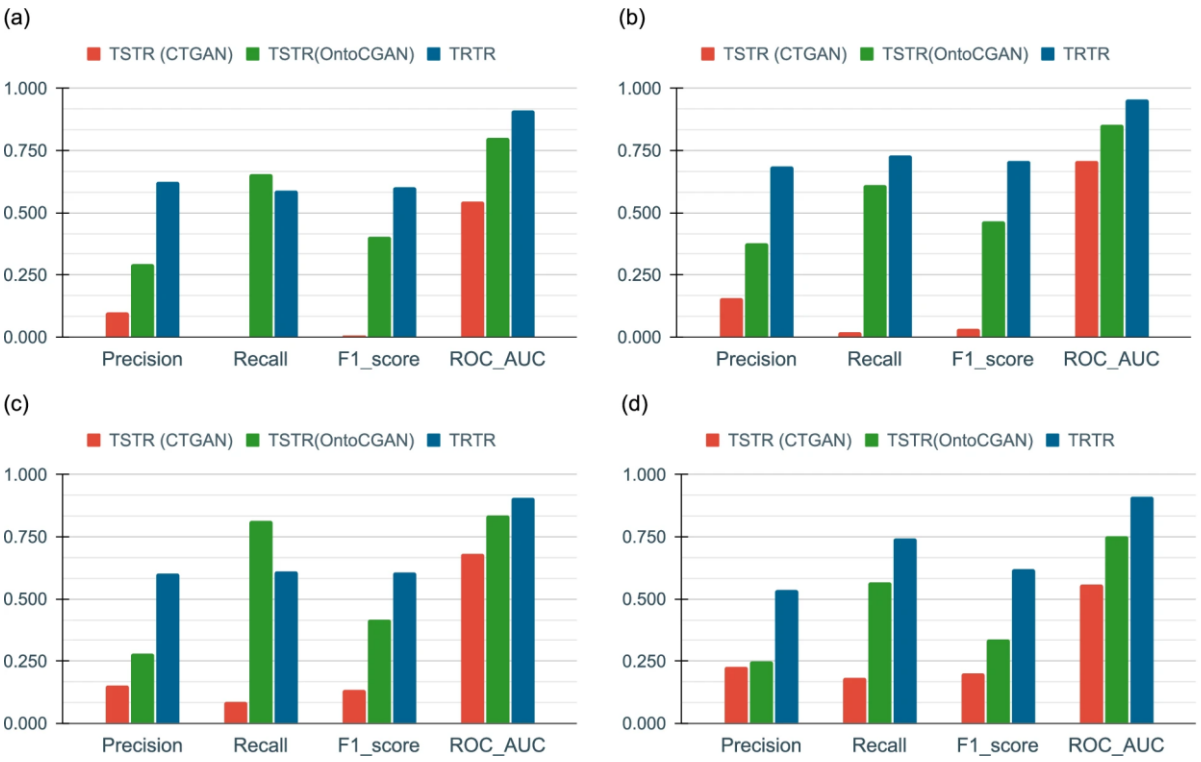

Figure 49. Machine Learning model performance using synthetic unseen data generated by Onto-CGAN, CTGAN, and real data.

A set of experiments were conducted to evaluate Onto-CGAN using MIMIC EHR datasets, with Acute Myeloid Leukemia (AML) chosen as the primary unseen disease scenario. The results, shown in Figure 49, highlight the advantages of ontology-guided generation. Across multiple evaluation metrics—precision, recall, F1-score, and ROC-AUC—Onto-CGAN (green bars) consistently outperformed CTGAN

(red bars) in the unseen disease setting. While models trained on real data (blue bars) remained the strongest overall, Onto-CGAN closed much of the gap and achieved substantial improvements in recall and ROC-AUC. In several cases, Onto-CGAN nearly doubled the recall of AML patient identification compared to CTGAN, while also producing more balanced precision–recall trade-offs. The F1-scores, although still lower than the real-data baseline, were significantly higher for Onto-CGAN than for CTGAN, confirming its ability to generate clinically useful training data in zero-shot scenarios.

We proved that Onto-CGAN is capable of producing synthetic records that not only reflect realistic distributions and comorbidity structures, but also support effective predictive modeling for previously unseen diseases. The integration of ontology-based embeddings ensures that generated cohorts align with established biomedical knowledge, allowing synthetic data to bridge gaps where real-world data is sparse or unavailable.

6 Cross-Modal Analysis: Achieving REALM's Integrated Framework Vision

With comprehensive synthetic data generation capabilities established across all major medical data modalities that directly support REALM's five use cases, we can now synthesize insights across our various approaches and show how our integrated framework exceeds the project's Beyond SoTA commitments while enabling the promised comprehensive evaluation capabilities.

6.1 Fulfilling the Comprehensive Generation Framework Promise

REALM Use Case	Modality Support	Project Commitment Fullfilled
Use Case 1 (Lung Cancer)	CT + Chest X-ray	medical image data generation
Use Case 2 (Pharmacogenomics)		
Use Case 3 (STAR Glucose)	Enhanced time-series	Demographic subgroup evaluation
Use Case 4 (COPD EHR model)	Privacy-preserving EHR	Formal differential privacy and rare cases generation
Use Case 4 (COPD time-series)	Enhanced time-series	Demographic subgroup evaluation
Use Case 5 (COPD outcomes)	Privacy-preserving EHR	Formal differential privacy and rare cases generation

Figure 50. Project Promise Fulfillment Matrix

A central research question underlying this section is whether multimodal synthetic data can in fact be generated in a coherent, clinically meaningful, and privacy-preserving manner. Addressing this question is essential for REALM, as the project vision explicitly requires an integrated framework that spans time-series, imaging (CT and chest X-ray), and structured EHR data. Only if synthetic data can be reliably created across these diverse modalities, and integrated into coherent patient profiles, can the promises of privacy-preserving clinical evaluation and digital twin simulations be fully realized.

The results presented in this section demonstrate that multimodal synthetic data generation is not only feasible, but can be achieved in a way that goes beyond the state of the art. Through innovations in distribution alignment, cross-domain transfer learning, ontology-guided generation, and formal

privacy integration, the consortium has established a comprehensive generation framework. The framework's outputs have been validated through multi-dimensional evaluation encompassing fidelity, utility, diversity, fairness, and privacy cost measurement, and have been shown to support clinical translation and regulatory compliance across all use cases.

Our work delivers the exact "comprehensive generation framework that combines data generators of various medical data types" promised in REALM Task 4.4, with each modality directly supporting specific use cases while contributing to an integrated evaluation ecosystem.

6.2 Beyond SoTA Achievements Validating Project Claims

The REALM framework has introduced a series of universal technical advances that go beyond existing synthetic data approaches. A key contribution is the innovation in distribution alignment, exemplified by the Enhanced TimeAutoDiff method, which addresses fundamental shortcomings in current generative techniques. Cross-domain transfer learning between chest X-ray and CT imaging has demonstrated novel capabilities in medical image synthesis, while ontology-guided generation has enabled rare disease modelling scenarios that were previously unattainable. Formal privacy integration represents another critical advancement, as it offers mathematically rigorous guarantees while preserving clinical utility. Moreover, the work demonstrated the conditioning effectiveness: All modalities benefited from explicit demographic conditioning, though the optimal conditioning strategy varied.
Another important aspect is the excellence of the evaluation framework developed in REALM. The project has delivered a multi-dimensional evaluation approach that assesses fidelity, utility, diversity, and fairness, and incorporates privacy cost measurements that allow informed trade-offs between privacy and utility. Subgroup evaluation capabilities using synthetic data provide more accurate and reliable possibilities for assessing fairness across demographic intersections. Together, these achievements form a comprehensive toolkit that fulfils the project's original promise of robust evaluation metrics and benchmarks for synthetic multimodal data.

6.3 Critical Insights Supporting REALM's Clinical Translation

The cross-modal analysis has yielded critical insights that directly support the clinical translation of REALM outputs. Our work demonstrates that synthetic data can provide more reliable subgroup evaluation than traditional approaches based on small real test sets, but this capability varies significantly across modalities and demographic groups. Modality-specific optimization has proven particularly effective: time-series data achieved the most consistent subgroup evaluation benefits, with success rates between 72 and 84 percent; Medical Imaging imaging, particularly CT scans, displayed superior fidelity and demographic conditioning compared to chest X-rays; EHR data successfully balanced privacy and utility through formal guarantees.

A critical finding across all modalities is the distinct challenges posed by evaluation utility (TRTS) compared to training utility (TSTR). The project has shown that generative models often produce smoothed versions of real data that are adequate for training but insufficient for evaluation. This gap, especially evident in time-series data, was addressed through explicit distribution alignment penalties in Enhanced TimeAutoDiff, which substantially improved evaluation performance. The recognition and resolution of this gap represent a breakthrough beyond the state of the art and directly enhance the reliability of REALM's evaluation framework.

Privacy-preserving capabilities have also been successfully demonstrated. Formal differential privacy mechanisms, inherent privacy in properly trained generative models that avoid memorization, and demographic conditioning without individual identifiers collectively ensure compliance with GDPR requirements while maintaining the clinical and statistical value of the generated data.

6.4 Platform Integration and Use-Case Support

The integration of synthetic data generation into the REALM platform architecture has been seamless. The AI orchestrator now supports modality-specific generation tailored to evaluation scenarios, provides quality-assured synthetic datasets validated through the comprehensive evaluation framework, manages privacy budgets for controlled data access, and enables real-time generation for dynamic use cases.

Each REALM use case has benefitted from targeted synthetic data support. Use Case 1 has leveraged advanced imaging with demographic conditioning for lung cancer evaluation. Use Case 3 has demonstrated enhanced time-series data with improved evaluation utility, while Use Cases 4 and 5 have relied on privacy-preserving EHR data that also facilitate rare phenotype modelling. In this way, cross-modal insights have strengthened each use case while ensuring consistency across the project.

6.5 Regulatory and Clinical Translation Impact

The REALM framework directly advances the evaluation of medical device software by enabling comprehensive bias assessments across demographic groups, ensuring privacy-preserving evaluation aligned with GDPR requirements, and supporting simulations of rare clinical scenarios that are otherwise unavailable.

From a regulatory science perspective, the formal evaluation frameworks and quantifiable privacy guarantees developed in REALM provide the methodological foundations necessary for regulatory submissions and medical device certification. These advances contribute not only to scientific progress but also to the project's commercialization objectives, by offering regulators and industry stakeholders tools that are both rigorous and trustworthy.

6.6 Limitations and Future Directions

While the results of cross-modal analysis are promising, several limitations remain. Clinical validation has so far involved limited direct input from medical experts, and further systematic involvement will be required to assess medical realism. The computational demands of comprehensive data generation are considerable, necessitating efficiency improvements for practical deployment. In addition, uncertainties remain regarding the generalization of models to datasets and populations not yet represented in the evaluation framework.

Future work will address these challenges by integrating clinical experts more systematically into validation processes, optimizing computational efficiency for scalability, and extending performance assessment to additional medical domains and diverse populations. These priorities will ensure that REALM not only continues to lead in synthetic data generation and evaluation but also achieves its long-term vision of providing an integrated, clinically validated, and regulatory-compliant framework.

6.7 Conclusion

Returning to the main question, the evidence gathered through T4.4 clearly shows that multimodal synthetic data can be generated effectively. Across time-series, imaging, and EHR data, we have demonstrated that synthetic datasets can reproduce key distributions, support subgroup-level fairness evaluation, and provide formal privacy guarantees. This supports REALM's broader objectives in digital medical device software evaluation and regulatory science. At the same time, the analysis has identified limitations, particularly the need for stronger clinical validation, efficiency improvements, and extended domain generalization, which will guide future development.

In conclusion, the cross-modal analysis confirms the feasibility and promise of multimodal synthetic data generation, while providing a realistic assessment of its current boundaries. It therefore directly answers the research question at the heart of REALM's integrated framework vision.

7 Synthetic Data Generation within the REALM architecture

Our comprehensive cross-modal analysis demonstrates successful fulfillment of REALM's integrated framework vision while providing honest assessment of current limitations. We now examine how these capabilities integrate into the REALM platform architecture to deliver the promised clinical evaluation and regulatory compliance capabilities.

7.1 Strategic Role of Synthetic Data in REALM Platform

The REALM platform represents a novel, comprehensive, and secure framework for medical device software evaluation, with synthetic data generation serving as a critical component within the Federated Cloud-based Data Repository . Our multimodal synthetic data generation capabilities integrate seamlessly into this architecture, providing essential functionality for scenarios where real-world data is insufficient, restricted, or fails to adequately represent diverse patient populations.

Strategic Role: The Data Synthesis component addresses fundamental limitations in traditional evaluation approaches by:

- **Augmenting scarce real-world data** for underrepresented subgroups identified through our comprehensive studies
- **Enabling privacy-preserving evaluation** through our differentially private EHR generation and inherently private image/time-series synthesis
- **Supporting comprehensive bias assessment** via our demonstrated capabilities in subgroup-level evaluation across multiple modalities

7.2 Platform Integration Overview

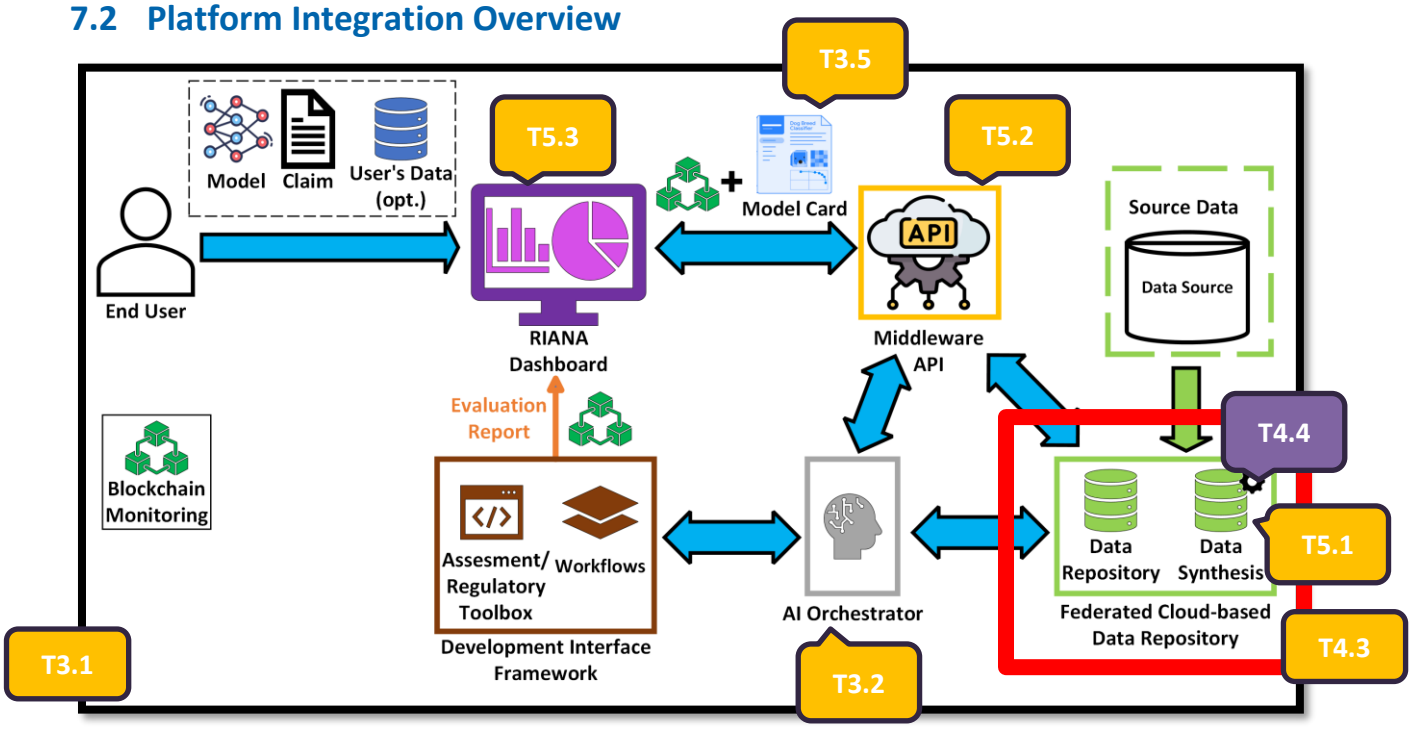

Figure 51 Conceptual architecture of the REALM platform- The Data Synthesis component, situated within the Federated Cloud-based data repository, is highlighted in red. Yellow boxes highlight the interaction of the work in T4.4 with other tasks.

Synthetic data generation as developed in Task 4.4 does not stand in isolation, but is conceived as an integral component of the overall REALM architecture. The methods, frameworks, and outputs described in this deliverable are designed to interact seamlessly with other tasks, ensuring that the project delivers a coherent and fully integrated platform for privacy-preserving medical data analysis and evaluation.

The first point of integration is with the REALM architecture defined in **Task 3.1**, where the system-level design provides the technical backbone that accommodates synthetic data generation and its downstream use.

In the backend of REALM exists an **AI orchestrator developed in Task 3.2,** which provides the decision-making and control logic within the platform. Synthetic data generation capabilities from Task 4.4 are directly orchestrated and managed within this environment, enabling scenario-specific data provision, privacy budget management, and real-time support for evaluation pipelines. The orchestrator constructs the pipeline comprising multiple steps necessary for assessing the model of interest.

A major dependency of the AI Orchestrator is the **Federated Cloud-based Data Repository established in Task 4.3** (outlined in red in Figure 51), which manages and provides functionalities related to evaluation data within the REALM platform. At a high level, it facilitates the storage of RWD for evaluations within REALM while also providing functionalities for synthetic data generation. This repository includes **Data Repositories** where RWD used for subsequent model evaluations is stored as well as **Model Repositories**, where the platform's internal models used for the generation and evaluation of synthetic data are saved.

Another important point of interaction is with the **RIANA dashboard developed in Task 5.3**, where synthetic data are made available in a user-accessible environment. The dashboard offers visualization and analytical tools that allow both technical and clinical users to interact with synthetic datasets and to evaluate their relevance and utility for specific use cases. Task 4.4 therefore underpins the dashboard with high-quality, privacy-preserving datasets.

Task 4.4 is also closely connected to the **data fusion and system integration activities of Task 5.2.** The communication between RIANA dashboard and other components is established through the middleware API. All information related to the evaluation process that has been entered by the user through the RIANA dashboard is encapsulated in a JSON file, which is sent to the AI orchestrator through the middleware API. This includes the configuration parameters for the synthetic data generator, if it is selected for use in the evaluation

When a new user first wants to interact with the REALM platform, the first step is to register through the **RIANA Dashboard developed in Task 5.3** The claim is then used to generate the **Model Card (T3.5)**, which acts as the structured repository of all necessary information required to initiate the creation of the evaluation pipeline within the REALM platform. This include the essential metadata about the model, as well as details on the datasets to be used, pre-processing requirements, and evaluation metrics. Additionally, based on the information in the Model Card, the AI Orchestrator determines which explainability techniques should be applied and whether **synthetic data generation** is required. These details allow the AI Orchestrator to configure the pipeline according to the specific validation task.

The first steps in the pipeline defined by the **AI Orchestrator** are primarily focused on preparing the necessary data for subsequent model evaluation steps. When the AI Orchestrator determines that synthetic data generation is required based on **Model Card** specifications, the **Data Synthesis** module activates appropriate generation capabilities:

1. **Modality-Specific Selection:** Based on the modality, the system selects the appropriate generation approaches:
   - **Enhanced TimeAutoDiff** for time-series ICU data requiring subgroup evaluation
   - **Fine-tuned latent diffusion models** for medical imaging with demographic conditioning
   - **DP-CGANS or Onto-CGAN** for EHR data based on privacy or rare disease requirements
2. The specific data and cohort to be generated is indicated in the Model Card, where the information about the data that the model under evaluation consumes or is expected to be evaluated against is included.

These initial steps, such as retrieving relevant data and generating synthetic data, ensure that the data required for model evaluation is available and properly formatted for the evaluation process. Once this data is prepared, it is passed to the next stages in the pipeline that focus on model evaluation.

T4.4 is highly linked with **Task 5.1**, which explores key issues in generating realistic and useful synthetic medical data, extending the approaches developed in T4.4. While Task 4.4 provides the generative frameworks, evaluation methods, and privacy-preserving mechanisms, Task 5.1 focus on post-market evaluation of the synthetic data created by previous tasks by Leveraging domain experts' insights, implementing advanced statistical methods for precise evaluation, and conducting adversarial evaluations between Real-World Data (RWD) and synthetic data using State-Of-The-Art (SOTA) models; this includes examining in-class and subpopulation performance, as well as assessing fairness and uncovering potential biases.

7.3 COPD Test Case

In this section, a complete example of the AI Orchestrator workflow is presented through a test case for a COPD dataset on a local node, excluding any federated aspects of the project. The workflow illustrates the implementation of the REALM platform’s architecture and the incorporation of the data architecture components. This section outlines the workflow initiated upon receiving evaluation instructions from the RIANA dashboard via the Middleware API, visually shown in Figure 52.

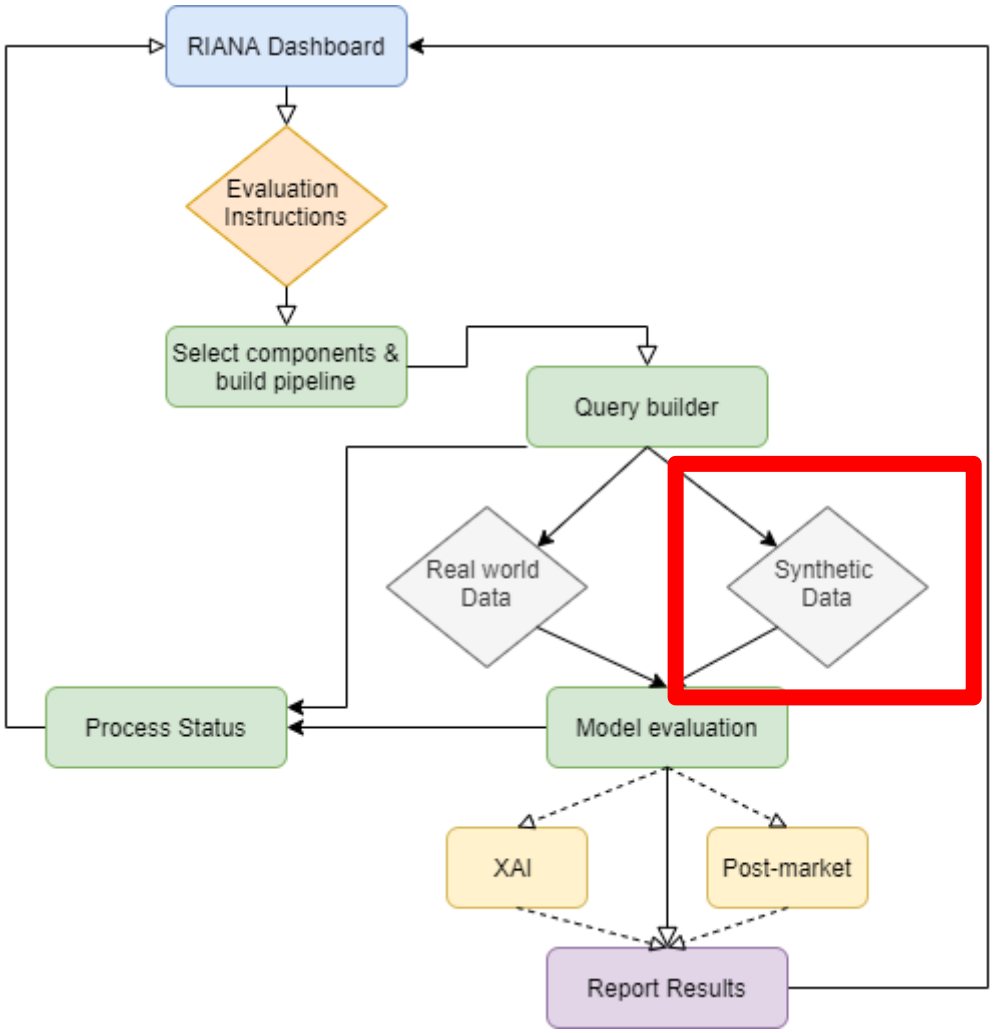

Figure 52 Flowchart of steps in AI Orchestrator workflow. The synthetic data component is outlined in red.

A major highlight of the workflow is the synthetic data generation component, which plays a pivotal role in enabling robust and inclusive model evaluations. Upon receiving evaluation instructions from RIANA, the AI Orchestrator interprets details such as the desired pipeline, data features, evaluation metrics, and requirements for synthetic data.

When synthetic data is required (either to compensate for limited real-world data or to simulate diverse evaluation scenarios), the workflow activates the synthetic generator component. If a pre-trained synthetic generator suitable for the user's needs is available on the REALM node, it is selected and employed to generate the necessary data. In cases where no matching generator exists, a new one is trained using the available data, ensuring that the synthetic dataset aligns with specific evaluation requirements.

By integrating the synthetic data generation module directly into the data loading sequence, the platform guarantees that model evaluation is not hindered by data scarcity or lack of representation. This approach ensures that the prepared data, whether real-world, synthetic, or a combination of both, is always available, realistic, and relevant for thorough and fair model assessment.

A pipeline example of the workflow described with a synthetic generator as loader can be found in Figure 53.

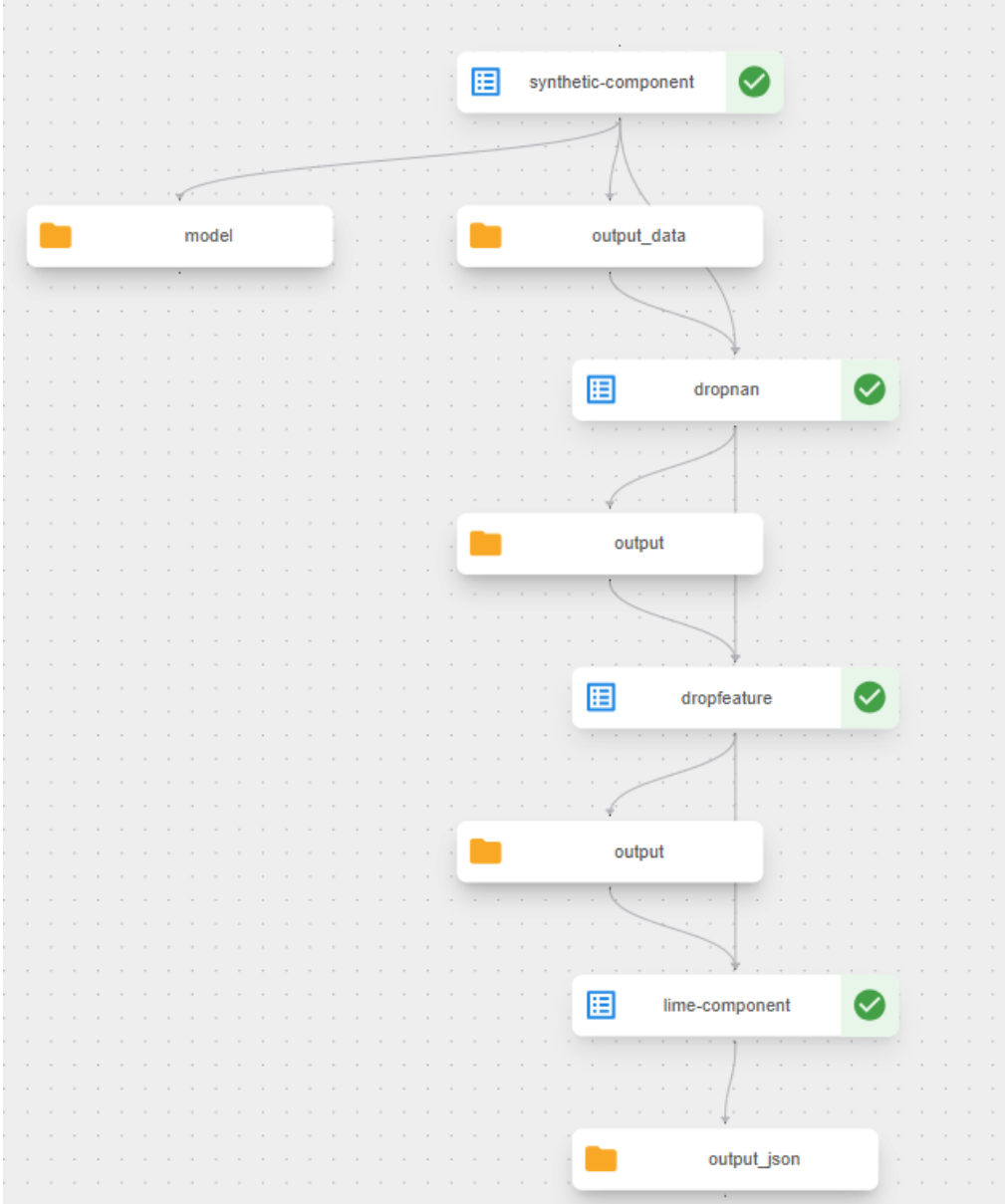

Figure 53 COPD with Synthetic generator pipeline

8 Open-source Contribution and Dissemination Activities within Task 4.4

Several dissemination activities were undertaken as part of Task 4.4 to share the outcomes, methodologies, and innovations with the broader research and clinical communities:

- **Peer-Reviewed Publications:**
  - Ibrahim, M., Al Khalil, Y., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., and Dumontier, M. (2025). "Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges," Computers in Biology and Medicine, 189, p.109834.
  - Sun, C. and Dumontier, M. (2025). "Generating unseen diseases patient data using ontology enhanced generative adversarial networks," npj Digital Medicine, 8(1), p.4.
  - Ibrahim, M., Sun, C., Elen, B., Ertaylan, G., and Dumontier, M. (2025). "Enabling Granular Subgroup Level Model Evaluations by Generating Synthetic Medical Time Series."
- **Conference Posters and Presentations:**

  - Ibrahim, M., Sun, C., Elen, B., Ertaylan, G., and Dumontier, M. (2025). "Enhancing AI model evaluation in medical applications using synthetic data," Poster presented at the MENAML conference.
  - Ibrahim, M., Sun, C., Elena, B., Ertaylan, G., and Dumontier, M. (2025). "Fine-grained performance evaluations using RW and synthetic data, and automated medical AI performance monitoring," Poster presented at the PhD Forum, ECML.
- **Keynote Presentation:**
  - Keynote at the ISC High Performance 2025 Synthetic Data event, Hamburg, highlighting advances and results in synthetic data generation and its applications.

Open Source Code Release:
The codebase for generative methods and evaluation frameworks developed in Task 4.4 was released as open-source, supporting transparency, reproducibility, and community engagement. The resources are available online at https://github.com/mahmoudibrahim98/realm_synthetic_data .

These dissemination efforts have been critical in ensuring that the findings and tools developed in Task 4.4 reach a wide audience of researchers, practitioners, and stakeholders, fostering further innovation and adoption in the field of synthetic medical data generation.

9 Conclusion

This deliverable, D4.6 "Multimodal Synthetic Data Generator", has reported on the work carried out in Task 4.4 between months 5 and 35 of the REALM project. The task was defined with the main objectives: (i) to generate synthetic multimodal medical data and complete patient personal profiles, (ii) to improve accessibility to personal medical data with privacy guaranteed.

The work presented here has demonstrated how each of these objectives has been fulfilled. Synthetic datasets have been successfully generated across multiple modalities, including time-series, CT, chest X-ray, and electronic health records, thereby addressing the first objective. These achievements directly respond to the central research question underpinning this work: can multimodal synthetic data be generated? The evidence presented here demonstrates that the answer is affirmative.
Privacy-preserving mechanisms, in particular differential privacy, combined with careful evaluation of privacy-utility trade-offs, ensure that data access is possible without compromising confidentiality, thereby fulfilling the second objective.
Crucially, these capabilities have been **operationalized to solve a persistent barrier in clinical AI evaluation: the lack of sufficient real-world data for minority subgroups**. By generating targeted, differentially private samples that reflect the relevant clinical and demographic characteristics, Task 4.4 enables **accurate, reproducible subgroup evaluation**, thereby aligning technical innovation with fairness and regulatory expectations.

In this way, Deliverable D4.6 provides both the methodological innovations, and the practical outputs needed to establish synthetic data generation as major pillars of the REALM framework.

9.1 Key Contributions and Achievements

The framework developed in Task 4.4 combines state-of-the-art generative methods with privacy-preserving mechanisms, resulting in datasets that capture the diversity and complexity of real-world clinical data without compromising patient confidentiality. Through innovations such as Enhanced TimeAutoDiff for **distribution alignment**, **cross-domain transfer learning for medical imaging**, **ontology-guided rare disease modelling**, and **differential privacy integration**, T4.4 has advanced the field of synthetic data generation beyond the current state of the art.

Among the principal contributions of Task 4.4 is the introduction of synthetic data–enabled subgroup evaluation as a systematic component of the REALM evaluation methodology. Conditional generators produce samples for specific intersections where real data are sparse, yielding evaluation sets with adequate coverage and statistical power across modalities. These synthetic test suites are calibrated against real hold-outs, assessed with multi-dimensional metrics (fidelity, utility, diversity, fairness, privacy), and governed by explicit privacy budgets. The result is a robust procedure that delivers trustworthy, fine-grained performance estimates for clinically salient subgroups, an advance that would be impractical using real data alone.

The outputs of Task 4.4 are tightly integrated with the broader REALM project. They provide foundational resources for Task 5.1 on synthetic data evaluation, fit perfectly in the REALM architecture (T3.1), RIANA Dashboard (T5.3), AI Orchestrator (T3.2), Model Cards (T3.5),Middleware API (T5.2), and are made available through the REALM data repository (T4.3) in compliance with FAIR principles. Importantly, the synthetic data developed here directly support the REALM use cases in Work Package 6, addressing needs in lung cancer imaging, glucose control, COPD management, and patient-reported outcomes.

As promised, Task 4.4 has also made tangible contributions in the form of **open-source code and publications.** These outputs are fully aligned with FAIR principles and ensure that the work can be reused and extended both within and beyond the project.

9.2 Impact on Healthcare AI Development

The outcomes of Task 4.4 have significant implications for the development and deployment of healthcare AI. By enabling the generation of multimodal synthetic data, the task provides a powerful means to overcome data scarcity, reduce reliance on sensitive patient information, and create robust training and evaluation pipelines for AI models.
By making subgroup evaluation accurate, scalable, and privacy-preserving, Task 4.4 directly strengthens the **safety and equity** of healthcare AI. Developers can identify performance gaps in under-represented populations early, quantify them with **tight uncertainty bounds**, and iterate with targeted data generation and model refinement. For regulators and clinical stakeholders, the approach provides **transparent evidence** that model behavior has been stress-tested across relevant subpopulations using methods that protect patient privacy. This shifts synthetic data from a convenience for development to a **cornerstone of reliable, fair, and compliant evaluation**, accelerating clinical translation while reducing the risk of harm to minority groups.

Moreover, the integration of privacy-preserving methods and fairness assessments ensures that the resulting datasets not only drive technical innovation but also align with societal, ethical, and regulatory expectations. This combination of utility, trustworthiness, and compliance represents a decisive step towards making synthetic data a mainstream enabler of healthcare AI. The contributions of Deliverable D4.6 thus go beyond technical progress: they establish a framework for safe, scalable, and clinically meaningful AI development, directly supporting REALM's vision of trustworthy, integrated, and patient-centered digital health solutions.

At the same time, the deliverable has provided a balanced assessment of current limitations. The absence of extensive clinical expert validation, the significant computational demands of large-scale synthetic data generation, and uncertainties regarding generalization across broader medical domains highlight areas where further work is needed. Addressing these gaps will require deeper integration of medical expertise, efficiency improvements in model architectures, and systematic validation across additional datasets and populations.

In conclusion, this deliverable confirms that multimodal synthetic data generation is both feasible and impactful. The results advances the state of the art and show that synthetic data can faithfully reflect real clinical information while preserving privacy. By delivering open-source tools, evaluation frameworks, and a scientific report, Task 4.4 has fulfilled its objectives and contributed significantly to the REALM promise of enabling privacy-preserving, trustworthy, and clinically relevant evaluation of medical device software.

10 Next steps

The next phase of work will focus on the complete integration of the synthetic data generation capabilities developed in Task 4.4 into the REALM production environment. This integration will take place within the framework of **Work Package 6**, where all REALM components are brought together to form the comprehensive REALM evaluation platform. The platform will provide the environment in which the five REALM use cases can be evaluated in a rigorous, standardized, and privacy-preserving manner. By combining synthetic data generation with data fusion, orchestration, visualization, and regulatory evaluation tools, the REALM evaluation platform will enable high-quality assessments of medical device software across diverse clinical scenarios. This step marks the transition from development to system-level validation, ensuring that the innovations of Task 4.4 achieve their intended impact in real-world healthcare AI applications.

11 References

Alaa, A. M., van Breugel, B., Saveliev, E., & van der Schaar, M. (2022). How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. *Arxiv*.

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, *28*, 525–552.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *International Conference on Machine Learning*, 214–223.

Bing, S., Dittadi, A., Bauer, S., & Schwab, P. (2022). Conditional generation of medical time series for extrapolation to underrepresented populations. *PLOS Digital Health*, *1*(7), e0000074. https://doi.org/10.1371/journal.pdig.0000074

Cao, R., Liu, Y., Wen, X., Liao, C., Wang, X., Gao, Y., & Tan, T. (2024). Reinvestigating the performance of artificial intelligence classification algorithms on COVID-19 X-Ray and CT images. *IScience*, *27*(5), 109712. https://doi.org/https://doi.org/10.1016/j.isci.2024.109712

Chambon, P., Bluethgen, C., Delbrouck, J.-B., der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., & Chaudhari, A. (2022). Roentgen: vision-language foundation model for chest x-ray generation. *ArXiv Preprint ArXiv:2211.12737*.

Chen, S., Ma, K., & Zheng, Y. (2019). *Med3D: Transfer Learning for 3D Medical Image Analysis*. https://arxiv.org/abs/1904.00625

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Donahue, C., McAuley, J., & Puckette, M. (2018). Synthesizing audio with GANs. *ArXiv*.

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (Vol. 4978, pp. 1–19). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79228-4_1

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). *Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs*. https://arxiv.org/abs/1706.02633

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, *30*.

Gunraj, H., Sabri, A., Koff, D., & Wong, A. (2022). COVID-Net CT-2: Enhanced Deep Neural Networks for Detection of COVID-19 From Chest CT Images Through Bigger, More Diverse Learning. *Frontiers in Medicine*, *8*, 729287. https://doi.org/10.3389/fmed.2021.729287

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2018). *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. https://arxiv.org/abs/1706.08500

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Arxiv*.

Ho, J., & Salimans, T. (2022). *Classifier-Free Diffusion Guidance*. https://arxiv.org/abs/2207.12598

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely Connected Convolutional Networks*. https://arxiv.org/abs/1608.06993

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *CoRR*, *abs/1901.07031*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. *Arxiv*. https://doi.org/10.48550/arXiv.1611.07004

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. M. (2021). *MIMIC-IV (version 1.0)*. PhysioNet.

Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. https://arxiv.org/abs/1901.07042

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *ArXiv Preprint ArXiv:1710.10196*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P., & Welling, M. (2022). Auto-Encoding Variational Bayes. *Arxiv*.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., & Lehtinen, J. (2023). *The Role of ImageNet Classes in Fréchet Inception Distance*. https://arxiv.org/abs/2203.06026

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. *Arxiv*.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z. A., & Yang, Y. (2022). RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, *4*(5), e210315. https://doi.org/10.1148/ryai.210315

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *ArXiv Preprint ArXiv:1411.1784*.

Models, M. G. (2023). *Chest X-ray with Latent Diffusion Models*. https://github.com/Project-MONAI/GenerativeModels/tree/7428fce193771e9564f29b91d29e523dd1b6b4cd/model-zoo/models/cxr_image_synthesis_latent_diffusion_model

Montet, F., Pasquier, B., Wolf, B., & Hennebert, J. (2024). Enabling Diffusion Model for Conditioned Time Series Generation. *Engineering Proceedings*, *68*(1). https://doi.org/10.3390/engproc2024068025

Nikitin, A., Iannucci, L., & Kaski, S. (2023). TSGM: A flexible framework for generative modeling of synthetic time series. *ArXiv Preprint ArXiv:2305.11567*.

Oord, A. van den, Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *ArXiv Preprint ArXiv:1711.00937*.

Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2337–2346.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, *5*(1), 1–13.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. https://arxiv.org/abs/2103.00020

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv Preprint ArXiv:1511.06434*.

Raoof, I., & Gupta, M. K. (2021). A Study of Semi Supervised Based Approaches for Motor Imagery Signal Generation. *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1312–1316.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2020). Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Arxiv*.

Roberts, A., Chung, H. W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., Hawthorne, C., Lewkowycz, A., Salcianu, A., van Zee, M., Austin, J., Goodman, S., Soares, L. B., Hu, H., Tsvyashchenko, S., … Gesmundo, A. (2022). Scaling Up Models and Data with t5x and seqio. *Arxiv*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. https://arxiv.org/abs/2112.10752

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Arxiv*.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Arxiv*.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Arxiv*.

Song, J., Meng, C., & Ermon, S. (2022). Denoising Diffusion Implicit Models. *Arxiv*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). *Going Deeper with Convolutions*. https://arxiv.org/abs/1409.4842

Tian, M., Chen, B., Guo, A., Jiang, S., & Zhang, A. R. (2023). Fast and Reliable Generation of EHR Time Series via Diffusion Models. *Arxiv*.

van Breugel, B., Qian, Z., & van der Schaar, M. (2023). *Synthetic data, real errors: how (not) to publish and use synthetic data*. https://arxiv.org/abs/2305.09235

Van De Water, R., Schmidt, H., Elbers, P., Thoral, P., Arnrich, B., & Rockenschaub, P. (2023). Yet another ICU benchmark: A flexible multi-center framework for clinical ML. *ArXiv Preprint ArXiv:2306.05109*.

Varma, M., Kumar, A., van der Sluijs, R., Ostmeier, S., Blankemeier, L., Chambon, P., Bluethgen, C., Prince, J., Langlotz, C., & Chaudhari, A. (2025). *MedVAE: Efficient Automated Interpretation of Medical Images with Large-Scale Generalizable Autoencoders*. https://arxiv.org/abs/2502.14753

Wang, T., Lei, Y., Fu, Y., Wynne, J. F., Curran, W. J., Liu, T., & Yang, X. (2021). A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, *22*(1), 11–36.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *Arxiv*. https://doi.org/10.48550/arXiv.1711.11585

Wang, W., Luo, J., Yang, X., & Lin, H. (2015). Data Analysis of the Lung Imaging Database Consortium and Image Database Resource Initiative. *Academic Radiology*, *22*(4), 488–495. https://doi.org/https://doi.org/10.1016/j.acra.2014.12.004

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, *2*, 1398-1402 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292216

Warvito. (2023). *Latent Diffusion Models for Chest X-Ray Generation using MONAI Generative Models*. https://github.com/Warvito/generative_chestxray/tree/main

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and Evaluation of Privacy Preserving Synthetic Health Data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2019.12.136

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Arxiv*. https://doi.org/10.48550/arXiv.1703.10593

12 Annexes

12.1 Annex 1 – Curated digital twin models for personalised simulations

This section summarizes the digital twins and computational models considered for the EDITH-use cases (Table A1), as well as those submitted from the wider community (Table A2).

The models and simulation workflows described under the **An Ecosystem for Digital Twins in Healthcare** Coordinate & Support action (EDITH-CSA). The **EDITH-use cases** (listed in Table A1) have been **implemented** within the proof-of-concept model infrastructure for the **Virtual Human Twin (VHT)**. These Digital Twin (DT) models can be accessed through the project catalogue/repository and deployed using the necessary high-performance computing resources.

Table A1: EDITH- Use cases, noted in EDITH-CSA – An ecosystem for digital twins in health care. The summary of use cases reproduced from Deliverable D5.1– PoC repository with available resources (EDITH – GA No. 101083771).

Osteoporosis *Bologna Biomechanical Computed Tomography Digital Twin model*	**University of Bologna, Italy**
The **Bologna Biomechanical Computed Tomography (BBCT)** model serves as a **digital twin** for a patient's hip, specifically designed to **predict the acute risk of hip fracture (ARF0)** based on a CT scan. The BBCT model significantly advances fracture risk assessment beyond the current clinical standard, which relies on **areal Bone Mineral Density (aBMD)**. The digital twin workflow uses **subject-specific finite element (FE) models** that are generated directly from the patient's Quantitative CT tomography scan data. The model is subjected to numerous **Finite Element (FE) simulations** (28 replicas) on an **HPC cluster** to simulate side-fall conditions using a spectrum of possible loads. The final result is the calculation of the **Absolute Risk of Fracture (ARF0)** and **Minimum Side-fall Strength (MSF)**, which quantify the patient's fracture risk. This allows for a more personalized and detailed analysis of the hip's biomechanical strength, for assessment of immediate hip fracture risk than traditional methods.	

Intensive Care Unit *STAR Digital Twin for Glycaemic control in ICU*[1]	**University of Liege, Belgium**
The **STAR Digital Twin (DT)** is a **patient-specific, risk-based decision-support system** designed for the Intensive Care Unit (ICU). Its primary function is to address the common and costly complication of **hyperglycemia** by providing **optimal glycemic control**. The STAR Digital Twin (DT) workflow is a two-step process focused on personalized glycemic control simulation: 1. **Input Data**: The workflow requires **anonymized patient clinical data**, which is essential for identifying the patient's unique insulin sensitivity profiles. This data includes the patient's diabetes status, nutrition intake, and records of insulin and glucose measurements. 2. **Forward Simulation**: The personalized **insulin sensitivity profiles** obtained in the first step are used as input for **forward blood glucose simulations**. The user can implement and test any given treatment protocol. The output of the DT simulation is the predicted **evolution over time** of blood glucose, interstitial insulin, and plasma insulin levels under the tested protocol.	
Cardiovascular *Atrial Modelling Toolkit for Constructing Bilayer and Volumetric Atrial Models at Scale*[2]	**Queen Mary University of London, UK**
The digital twin within the **Atrial Modelling Toolkit**, an **open-source pipeline** designed to rapidly and reproducibly construct personalized **cardiac models of the atria** for large-scale *in silico* trials. It overcome the challenge of slow, non-reproducible cardiac model construction to enable **large *in silico* trials** and **personalized predictions** on clinical timescales. The DT workflow uses s**egmentation masks** of the atria, as derived from raw MRI or CT data, or provided directly by the user, as input. The underlying computational model generates a **simulation-grade atrial mesh** and corresponding initial conditions. This mesh incorporates critical anatomical features: **transmurability** (structure through the wall), different **atrial regions**, and **atrial fibers**. The digital twin facilitates a personalised **electrophysiological simulation,** using the input mesh and initial conditions. The DT simulation's output is used to investigate the effects of anatomical factors (like fibers and **fibrosis**) on the dynamics of **atrial fibrillation** (fibrillatory dynamics).	
Cancer *Cancer Diagnosis Based on omics Information - PhysiBoSS 2.0*[3]	**Barcelona Supercomputing Center, Spain**
The cancer digital twin workflow is a computational process that uses **HPC resources** and the **PhysiBoSS tool**[3] to simulate a personalized population of **tumour cells**. The digital twin model considers **patient data**, including clinical variables (e.g., age, disease status), **single-cell sequencing data** (CSV and hdf5 formats), and optionally **CT scans** (TIFF/DICOM) for tumour size and morphology estimation. The personalized Boolean model from single-cell sequencing and the initial spatial configuration derived from CT scans (when available) are fed into the **PhysiBoSS tool** to run the cell-level simulation. PhysiBoSS provides the final cell population, allowing for the tracking of each cell's **spatial location** and **phenotypic attributes**, along with intermediate snapshots to track evolution. The simulation results are used to generate an **HTML report** for non-advanced users, including	

relevant metrics like the final tumour cell count, which can help infer tumour subtype or aggressiveness.

The models listed in Table A2 are **external digital twins or computational models** that have been curated for the EDITH Proof-of-Concept model catalogue. These models are **not yet implemented**; instead, they were selected by the community for their potential use within the future **Virtual Human Twin (VHT) infrastructure**.

Table A2: Digital twin models curated from the community, by EDITH-CSA. The table is reproduced from EDITH-CSA – An ecosystem for digital twins in health care. Deliverable D5.1– PoC repository with available resources (EDITH – GA No. 101083771).

S.NO	Medical domain, Pathology, organ systems	Description of the Digital Twin and/or underlying computational model	Organisation, Country	Publication, model source
1	Rheumatoid arthritis	Rheumatoid arthritis model	University of Toulouse III-Paul Sabatier, France	REF1; REF2; REF3
2	Orthopaedics	Gene/protein regulatory network model for identifying therapeutic targets against osteoarthritis.	KU Leuven, Belgium	REF1; REF2
3	Pharmacokinetics	Digital twin for Immune system - modelling toolbox for drug development .	Sanofi, Belgium	
4	Cardiovascular	Simulation of patients with aortic aneurysm.	Università di Palermo, Italy	REF1
5	Respiratory	Lung poromechanical model.	Ecole Polytechnique, France	REF1; REF2
6	COPD	A simulation model for Chronic Obstructive Pulmonary Disease from Lung CT, integrating inflammatory mediated remodelling of lung tissue and consisting of 2 weakly coupled models: an ABM based for the inflammation part of the immune system and a FEM model to calculate strain due to cyclic respiration.	Joint Research Center (JRC), Europe	REF1
7	Imaging	Whole body CT-based digital twin, to monitor the progression of organs and tissues as a function of disease and interventions, and build predictive models for clinical outcomes.	Voronoi Health Analytics Inc. Canada	Website
8	Imaging	CT whole body automated segmentation.	Voronoi Health Analytics Inc. Canada	Website
9	Psychiatry	Digital twin of patients with mental illness (VirtuVP), that simulates realistic psychiatric interviews while automatically assessing the communication skills of students and clinical professionals in psychiatry.	Laboratory of Immersive Neurotechnologies, Spain	REF1 REF2
10	Surgery	Model and methodology to compute forces during surgeries that surgeons experience when operating.	University of Bremen, Germany	REF1

11	Musculoskeletal	Musculoskeletal motor control	Carnegie Mel-lon University , USA	REF1
12	Pharmacokinetics	PBPK model predicting post-hepatectomy survival	Humboldt-University Berlin, Faculty of Life Sciences, Germany	REF1; REF2
13	Musculoskeletal	Articulated kinematically driven skeleton motion	German Cancer Research Center, Germany	REF1

[1] Lin, Jessica et al. "A physiological Intensive Control Insulin-Nutrition-Glucose (ICING) model validated in critically ill patients." *Computer methods and programs in biomedicine* vol. 102,2 (2011): 192-205. doi:10.1016/j.cmpb.2010.12.008

[2] https://github.com/pcmlab/atrialmtk

[3] Ponce-de-Leon, M., Montagud, A., Noël, V. *et al.* PhysiBoSS 2.0: a sustainable integration of stochastic Boolean and agent-based modelling frameworks. *npj Syst Biol Appl* **9**, 54 (2023). https://doi.org/10.1038/s41540-023-00314-4